学位論文 博士 (工学)

Web上のテキストデータを用いた道徳的常識の自動獲得に関する研究

2018年度

慶應義塾大学大学院理工学研究科

山本 眞大

目 次

図 目 次

表 目 次

第 1章

序論

1.1 はじめに

本研究の目的は，機械に道徳的常識を低コストで獲得させることである．本研究において道徳的な常識とは,「社会や共同体において，その構成員の大多数によって共有される道徳観に基づき，より健全で快適な共同生活を送る為に守るべき，又は行うべきと考えられている規範，行動の指針」と定義される [4]．つまり，ある行為の社会的善悪性を低コストで推定することを目的とする．

上述した知識を幅広くかつ低コストで獲得するために，Web 上に存在する大規模なテキストデータを用いる．Web 上には多種多様なデータが大量に存在するため，自然言語処理技術を適用することで，様々な知識を獲得できると考えられる．具体的には，(1) 評価表現を用いて道徳判断を行う手法 (第 2 章), (2) 分散表現を用いて道徳判断を行う手法 (第 3 章)，(3) 道徳的常識を自動獲得した上で，獲得した知識を深層学習を用いて学習し，道徳判断を行う手法 (第 4 章) を提案することによって上記の実現を目指す．

本研究は分野横断的な研究であるため，複数の学問領域と関係する．本章では，その中でも特に関わりの深い,「人工知能 (Artificial Intelligence, AI)」,「ロボット倫理学」,「自然言語処理」についてまとめた後，本研究の位置付けを述べる．

本章は以下のように構成される．まず，1.2 で人工知能の歴史を概略した後，近年顕在化してきた人工知能の道徳性の問題について述べる．1.3 では，人工知能の道徳性の問題について扱う学問であるロボット倫理学について説明する．1.4 では，テキストデータを処理する上で必要となる自然言語処理技術について概説する．1.5 で本研究の位置付けについて述べた後，1.6 で本論文の構成についてまとめる．

1.2 人工知能

本節では人工知能の歴史及び，近年顕在化してきた人工知能の道徳性の問題についてまとめる．

1.2.1 人工知能の歴史

1950〜1970 年代

1956 年のダートマス会議をきっかけとし第 1 次 AI ブームが到来した．人間の知能や知性を実現する「強い AI」が研究目標とされており，様々な研究が行われた．例えば，1966 年には対話システムの ELIZA が開発されている [5]．当初は，極めて楽観的な見通しであったが，結果的には単純なゲームや迷路の探索程度しかできなかった．特に，実世界の複雑な問題の解決に利用できないことが大きな問題であり，1969 年，ジョン・マッカーシーとパトリック・ヘイズにより，有限の情報処理能力しかないロボットには，現実に起こりうる問題全てに対処することができないとされる「フレーム問題」が指摘された [6]．

1980 年代〜2000 年代

コンピュータの性能が向上することに従い，人工知能研究にも 2 回目のブームが訪れた．この時代に行われたのは，知識やルールを人間が記述し，それらに基づいて特定の知識処理を行わせるという取り組みである．この取り組みは，特定分野の専門家の知識を記述するエキスパートシステムとして成果をあげることとなった．一方で，ルールを記述するのはあくまでも人間であり，世の中のあらゆる事象を記述することが困難であるという問題が存在した．その結果，汎用的なシステムの構築はほとんど行われないまま，ブームの終焉を迎えた．

2010 年代〜

ムーアの法則に従うコンピュータの性能向上及び，インターネットやスマートフォンの登場によるデータの爆発的な増大に伴い，第 3 次人工知能ブームが到来した．既

に特定分野においては人間の性能を上回る事例が多数報告されており，実用化されている技術も多数存在する．

2011 年には，米国のクイズ番組「Jopardy!」にて IBM の「Watson」がクイズチャンピオンに勝利した．2012 年には，画像認識のコンテスト ILSVRC（ImageNet Large Scale Visual Recognition Challenge）でカナダのトロント大学のチームが深層学習により 2 位と圧倒的な差をつける成果を残している [7]．

コンピュータ将棋の世界では，2010 年に「あから 2010」が清水女流王将を破ったことが報告されている [8]．その後，プロ棋士とコンピュータ将棋ソフトウェアが対戦を行う将棋電王戦がドワンゴにより開催された．情報処理学会は 2015 年 10 月に，コンピュータ将棋の実力は 2015 年時点でトッププロ棋士に追いついているという分析結果を出している [9]．またプロ棋士の間でも，2017 年には「Ponanza」が佐藤天彦名人を破ったことなどを受け，人間を超えたとの認識が広まっている．

コンピュータ囲碁の世界では，AlphaGo [10] がイ・セドル九段に勝利し話題になった．その後同グループは，人間の棋譜を用いず，ルールのみを用いて強化学習を行うことで AlphaGo よりも強いプログラムの開発に成功したと報告している [11]．さらに同様のアルゴリズムを将棋やチェスに適用することにより，従来手法よりも強いプログラムの開発に成功したと発表している [12]．

自然言語を別の自然言語に翻訳する変換をコンピュータを利用して自動的に行う機械翻訳分野においても，深層学習によるブレークスルーが起きている．特に注意機構の導入に伴い，既存手法の精度を大きく上回る報告がされている [13,14]．2016 年には Google の翻訳システムも深層学習を用いたものに置き換わり，その高品質な翻訳の実現が世間に大きな衝撃を与えた [15]．

人間とインタラクションを行う対話システムの研究開発，実用化も 1 つのトレンドとなっている．2014 年には，ウクライナ在住の 13 歳の少年という設定の下，チューリングテストをクリアした対話システムが登場した [16]．対話システムは大きく分けて 3 つに分けることができる．1 つ目は，ある特定の目的を持つタスク指向型の対話システムである．例えば，コールセンターで応対を行うチャットボットなどがこの分類に相当する．2 つ目は，特定の目的を持たない非タスク指向型の対話システムである．例えば，Microsoft 社の「りんな」や「Tay」などがこの分類に相当する．3 つ目は，非タスク指

向とタスク指向の両方を含むものである．これは，対話システムの実用化により新たに生まれた分類である．例えば,「Yahoo!音声アシスト」では，天気や今日の予定を知る機能に加え，雑談を行う機能も搭載されている．その他，Apple の「Siri」，Google の「Google Assistant」などの音声アシストや,「Google Home」,「Amazon Echo」,「LINE Clova WAVE」などのスマートスピーカーも 3 つ目の分類に相当すると考えられる．

自動運転に関する研究開発及び実用化も数多くなされている．自動運転技術はレベル 1 (運転支援) からレベル 5 (完全自動運転) までのレベルに定義されており，現在ではレベル 2 の部分自動運転が実用化されている [17].

上記のように様々な場面で人工知能技術が有用であると認識されている一方で，人工知能に関する道徳的な問題が顕在化してきている．次項では，その代表的な事例について述べる．

1.2.2 人工知能と道徳性

本項では，人工知能技術の道徳性について世間で物議を醸した事例について述べる．ここで述べる事例は，1.3 で説明するロボット倫理学という学問分野が議論の対象としている事例である．

自動運転技術による初めての死亡事故

2016 年 5 月 7 日，テスラ・モーターズカー社の自動運転技術を搭載した車「モデル S」により死亡事故が発生した [17]．自動運転車による初めての死亡事故であり，大きな物議を醸した．特に，自動運転車が事故を起こした場合，誰が責任を取るのかという問題が論じられた．

この自動運転車に関して倫理的な問題が必要となるのは，正常な動作の結果，不可避的に事故が生じる場面にある．例えば，このままの状態では事故を起こす場合，右にハンドルを切ればその事故が防げるとする．しかし，右にハンドルを切った場合，別の事故を起こす可能性がある．その場合，人工知能はどちらの事故を防ぐべきかを判断することになる．上記の問題はトロッコ問題と呼ばれる倫理学において有名な問題である [18]．このまま自動運転技術が高度化した場合，上記のような「人の命に定量

的な価値をつける」という行為が発生する可能性が高く，このような行為は道徳的に容認されるべきかという倫理的な問題が生じる．

爆弾ロボットによる犯罪者の殺害

2016 年 7 月 7 日，米テキサス州ダラスで警察官狙撃事件が起きた．これに対しダラス市警は，殺人ロボットを投入して犯人を殺害した．リモコン操作による殺害のため，人間が間接的に手を下したことになるが，ロボットが殺人を犯した初めての事件として注目された [19]．今後，自立型のロボットが登場する可能性があるが，殺人を犯すことは正しいことなのか否かの判断を如何に実装するかの問題が残る．それと同時に，その責任を取る人は誰になるのかといった問題や，そもそもそのような価値判断を行わせることが正しいのかといった倫理的な問題も生じる．

黒人をゴリラと誤認識する Google の画像認識 API

2015 年 5 月 29 日にリリースされた Google の写真アプリ「Google Photos」には，アップロードされた写真に写っている物に対して自動的にタグ付けを行う機能がある．その機能によって，黒人 2 人組が映る写真に「ゴリラ」というタグが付与される事件が起きた．人種差別に繋がる事件であり，世間の注目を集めた [20]．

Microsoft 社の対話ボット Tay

Microsoft 社の対話ボット Tay は 2016 年 3 月 24 日，SNS 上で非道徳的な発言を繰り返し炎上した [21]．この対話ボットは，ユーザとのインタラクションの結果，新しい表現を獲得していくというものである．そのため，悪意あるユーザが非倫理的な発言を繰り返すことにより，その発言を覚えてしまうということが生じた．

人間と機械の間のインタラクションの機会は，今後ますます増加していくと考えられるが，このような事件を未然に防ぐためにも，人工知能に道徳的な知識を獲得させることは極めて重要であると考えられる．

まとめ

本項では，人工知能の実用化にあたり，倫理的な側面について世間で話題になった事例について述べた．上述したように人工知能技術の発展により我々の生活が便利になる一方で，その倫理的な問題が認知されるようになってきた．次節にて，人工知能と倫理に関する議論を総合的に扱う学問分野，ロボット倫理学について概説する．

1.3 ロボット倫理学

ロボット倫理学とは，ロボットに関する倫理的問題を扱う応用倫理学の一分野である．その対象の範囲は多岐に渡り，例えば以下のような事例が議論の対象とされている [22].

- データを収集する際のプライバシーの問題
- 戦争における無人機や自律型兵器の是非
- コンパニオンロボットが人間の心理や人間同士の関係に与える影響
- ロボットの倫理的判断の問題

1.3.1 ロボット倫理学に関する取り組み

本項では，人工知能の倫理的な問題に対する各国の取り組みについてまとめる．人工知能によって可能なことが増加するに従い，各国でロボット倫理について考えられるようになった．

人工知能学会の倫理指針

日本では，人工知能学会の倫理委員会によって作成された「人工知能学会 倫理指針」が，2017 年 2 月 28 日の人工知能学会理事会によって承認されている [1,23]．A.1 に人工知能学会の倫理指針の原文を示す．この倫理指針は全部で 9 つの項から成り，主に人工知能学会会員の倫理的な価値判断の基礎となる倫理指針について定められている．例えば，他者のプライバシーの尊重や人工知能の安全性，社会に対する責任などの項

が存在する．他と比べて特徴的なものは 9 項目の人工知能への倫理遵守の要請である．この項では,「人工知能が社会の構成員またはそれに準じるものとなるためには，上に定めた人工知能学会員と同等に倫理指針を遵守できなければならない」と定義されており，自律的な人工知能にもこの倫理指針が適用されるという再帰性を含んでいる．

内閣府による「人工知能と人間社会に関する懇談会」

内閣府では，2016 年 5 月に「人工知能と人間社会に関する懇談会」が設立された [24]．倫理的な観点からは，意識や心を持つ人工知能の扱いや，人間の意思決定や信念の形成に関する人工知能の扱いに関して検討されている．

アシロマ会議

2017 年のアシロマ会議では人工知能に関する専門家が集まって議論を行い，人工知能に関する 23 の原則が発表された [2]．A.2 に「アシロマ AI 23 の原則」を示す．この発表された内容については，2017 年 9 月の段階で約 1200 人の人工知能研究者がサインしている．特に価値観と調和に関する原則では,「高度な自律的な人工知能システムは，その目的と振る舞いが確実に人間の価値観と調和すべき」と定められている．この目標を達成するためには，人間が当たり前のように保持している行動に関する価値観，つまり本研究で対象とするような道徳的常識が必要であると考えられる．

その他の活動

韓国では，2007 年に「ロボット倫理憲章」の草案が発表された [25]．全 7 条から成り，主に人間中心の倫理規範について述べられている．国際的には，Future of Life Institute (FLI), Machine Intelligence Research Institute (MIRI), OpenAI などの様々な組織が，人工知能の倫理的な側面についての議論を行っている．

国際会議や大学の講義でもロボットの倫理的問題について取り上げられることが増えてきている．国際会議では, International Conference on Robot Ethics and Safety Standards 2017 (ICRESS 2017) のようなロボット倫理学を主体とした会議が開催されている．また自然言語処理系の国際会議では “Ethics in Natural language processing”

というワークショップが開催され, 2017 年に引き続き 2018 年も開催される予定となっている.

大学の講義でも, 倫理的側面を主題としたものが増えてきている. 2017 年にはコーネル大学にて “Ethics and Policy in Data Science” という授業が, ワシントン大学にて “Ethics in NLP” という授業が開催されている. 2018 年にはカーネギーメロン大学にて “Computational Ethics for NLP” という授業が開催されている.

1.3.2 ロボット倫理学の研究例

本項ではロボット倫理学に関連する研究についてまとめる. 中でも特に本研究と関わりの深い, ロボットに道徳性を習得させることを目的とした研究について概説する. 大きく分けて以下の 2 つの研究が存在する.

理論的な研究

1 つ目の研究は, 理論的な研究である. 論理学を基に提案されているものが多い. 例えば, Bringsjord らは義務論理 [26], Powers らは非単調論理 [27], Pereira らは予測論理 [28] を基にした研究を発表している. しかし, これらの研究は提案のみに留まっており, 実装・評価は行われていない.

実装・評価が行われている研究

Anderson らは, MedEthEx というシステムを提案している [29]. これは, ロボット NAO に医療分野において善悪判断を行うシステムを組み込んだものである. 具体的には, 患者が薬を飲むことを拒否したときに, 患者の意思を尊重するのか, それとも患者の健康を重んじて担当医に連絡をするのかの判断を行うシステムを提案した.

McLaren は Truth-Teller および SIROCCO というシステムの構築を行った [30, 31]. Truth-Teller はエキスパートシステムの一種で, ユーザの状況に応じて本当のことを言った方が良いか, 黙る方が良いかなどの計算を行うシステムである. SIROCCO は, 法律分野において倫理的な判断を行うエージェントである. 法律分野の判例に基づく

推論を行うことで，技術者が仕事で直面するジレンマについてアドバイスをするプログラムである．

以上，実際に実装・評価が行われている研究について述べたが，これらの研究には以下に述べる 2 つの問題が存在する．

- 知識の網羅性の低さ: いずれの手法も特定のドメインを対象としているため，我々が一般的に保有している常識については考慮されていない．
- システムを実装する際のコストの大きさ: 人手によるルールや専門家による大量の推論事例が必要であり，システム実装のためのコストが大きい．

1.4 自然言語処理

本節では，道徳的な常識をテキストデータから獲得するにあたり，関連する自然言語処理の技術について述べる．

1.4.1 知識獲得

自然言語処理の究極の目標は，言語を理解する機械を作ることである．この目標に向けて，我々人間が暗黙のうちに共有している常識を，機械にも習得させる研究が存在する．知識獲得研究は，単語を対象とする研究と，単語よりも大きな言葉の単位を対象とする研究に分けることができる．

単語を対象とする研究

単語を対象とした知識獲得の研究は幅広く，例えば以下のような知識を収集する研究が存在する．同義語・類義語獲得研究では，その名の通り,「本」と「書物」のような同じ意味を持つ語の獲得を目的とする．これらの知識は，国語辞典を用いた手法 [32] やパターン情報を用いた手法 [33]，ニューラルネットワークにより学習された単語ベクトルを用いた手法 [34,35] などにより獲得できることが示されている．また，これらの同義語情報は，検索エンジンのクエリ拡張などに用いられている [36].

関係情報や属性情報について扱う研究も存在する．関係情報としては,「本」と「教科書」や「楽器」と「ピアノ」などの上位下位関係などがあり，属性情報としては,「りんご」は「赤い」などの，物に対するサイズや色の情報などがある．関係抽出分野では，教師あり学習を用いた手法 [37–39]，教師なし学習を用いた手法 [40, 41] に加え，Distant Supervision と呼ばれる，コストをかけずに大量の擬似ラベル付きデータを生成して学習を行う手法も提案されている [42–46].

関係性を明確に定義せず，関係の度合いを定義し分類する研究も存在する．この研究分野では，あるキーワードが与えられたときにそのキーワードから想起されやすい語を関連語として収集することを目的としており，収集対象の知識は連想知識と呼ばれることもある．基本的には Web 上のテキスト中の単語間の共起情報を用いる研究が多い [47] が，最近では，ゲーミフィケーションを利用してこれらの知識を収集する研究も存在する [48, 49].

常識的知識を整理したデータベースには以下のようなものがある．WordNet [50, 51] は，同義関係にある単語を同一の synset にまとめ，その定義や他の集合との関連を人手で整備している．Cyc [52] は Lenet らによって開始された，常識的知識をデータベース化するプロジェクトである．約 50 万の単語が登録されており，人の持つ常識のうち約 2%程度の常識を蓄えているとされている．MIT の Open Mind Common Sense (OMCS) [53] では，一般のボランティアから常識的知識を獲得するアプローチを取っている．獲得された知識は ConceptNet [54] と呼ばれるネットワークに登録されている．

日本語のデータベースも多岐に渡り，例えば，日本語 WordNet [55]，EDR 電子化辞書 [56]，日本語語彙大系 [57, 58]，連想概念辞書 [59]，日本語 Wikipedia オントロジー [60] などが存在する．

単語よりも大きな単位を対象とする研究

単語よりも大きな単位を扱う研究も存在する．例えば事態間知識は,「犯罪を犯すと，警察に捕まる.」というような 2 つの事象間の関係に関する知識を対象とする．テキストデータから事態間関係知識を獲得する研究には，述語項構造の分布類似度を用いるもの [61]，共起パターンを用いるもの [62]，両者を併用するもの [63] などが存在する．またここで獲得された知識は，Winograd Schema Challenge (WSC) [64] と呼ばれる，

常識的知識を必要とする共参照解析タスクなどに適用されている [65–67].

スクリプト知識は，ある状況において典型的に起こる一連の出来事を記述した知識であり，1977 年に Schank らによって提唱された [68]．例えばレストランに関するスクリプトは,「レストランに入店する」→「メニューを注文する」→「食事する」→「代金を支払う」→「レストランを出る」のように定義できる．このような知識は，人工知能が適切な状況の理解を行うために必要なものであると言われている．しかしながら，ありとあらゆる状況についてこのような知識を人手で書き下すことは困難であるため，自動でスクリプト知識を獲得する研究が行われている [69, 70].

1.4.2 Sentiment Analysis

Sentiment Analysis とは，ある対象に対する個人の評価情報を分析する分野である [71]．例えば，製品やサービスなどの世評の解析がこの分野の研究例としてあげることができる．2004 年春に人工知能研究に関する国際会議 AAAI で最初の会議が開催されて以来，様々な国際会議のワークショップやシェアドタスクでの研究テーマとなっている．

日本語のデータセットも多数構築されている．単語レベルで評価極性が付与されたデータセットとして，小林らの評価値表現辞書 [72, 73]，高村らの単語感情極性対応表 [74, 75]，鍛冶らの Polar Phrase Dictionary [76]，佐野らのアプレイザル評価表現辞書 (JAppraisal 辞書) [77] などが存在する．文や文章レベルで評価極性が付与されたデータセットとして，筑波大学文単位評価極性タグ付きコーパス [78]，ACP コーパス [79, 80]，京都観光ブログの評価情報付与データ [81] などが存在する．

Sentiment Analysis 分野には，商品のレビューが肯定的なのか否定的なのかを判断する Sentiment Classification タスク，評判情報を構造化するタスク，評判箇所を検出するタスクなど，様々なタスクが存在する．以下本節では，本研究とタスク設定が類似している Sentiment Classification タスクにおける既存手法を 4 つの観点からまとめる．

教師なしの手法

まず初めに，ラベルデータが全く与えられていない場合の手法について述べる．最初に提案されたのは，Turney らの手法である [82]．アイデアは，対象とする文と excellent や poor などの評価表現との共起情報を用いるものである．例えば，検索エンジンを用いて上記クエリのヒット件数の大小によって対象の極性を推定する．

半教師あり学習に基づく手法

次に，半教師あり学習に基づく手法について説明する．半教師あり学習では，少量のラベル付きデータと大量のラベルなしデータを使用して分類モデルの構築が行われる．半教師あり学習は,「ラベル付きデータを用意することはコストがかかるが，ラベルなしデータは大量に用意することができる」という現実的な問題設定であるため，近年盛んに研究が行われている [83–87]．

教師あり学習に基づく手法

3 つ目に，あらかじめラベル付きデータが十分に与えられている場合の手法について概説する．Pang らは，評判分析を始めて機械学習の問題として定式化した [88]．その後，5 段階のレーティング予測 [89] などが行われている．

教師あり学習に基づく手法の焦点は，2 つある．1 つ目は特徴量抽出の問題である．一般に，入力は文または文書などの言語データであり離散的である．そのため，コンピュータが処理可能な数値情報に変換する必要がある．特徴量抽出の際には，単語 uni-gram や単語 bi-gram を特徴量として用いる方法が一般的であるが，それ以外の情報を用いる研究も存在する．例えば，Matsumoto らは，語の系列や語間の依存構造情報などの統語情報を特徴量として使用する手法を提案している [90]．

もう 1 つの問題は，モデル選択の問題である．Sentiment Classification は，機械学習分野で一般的な分類の問題として定式化可能であるため，そのモデルの候補は少なくない．例えば，ナイーブベイズ分類器や，最大エントロピー法に基づく分類器，Support Vector Machine (SVM) による分類器などが使用されることがある．

一方近年では，深層学習を用いたモデルも多数提案されている．Sentiment Classification 分野において深層学習を用いた初期の研究は，再帰型ニューラルネットワークを用いたものが多い．例えば，Recursive Neural Network [91, 92] を用いた研究や Recursive Tensor Network [93] を用いた研究などが挙げられる．機械翻訳や自動要約で高精度を記録した時系列ニューラルネットワークを用いた研究も存在する．代表的なものは，Recurrent Neural Network を用いる研究 [94] や Long Short-Term Memory (LSTM) を用いる研究 [95] などである．また，言語の表層的な情報だけではなく，依存構造情報などを用いる研究として，Tree-LSTMs を用いる手法も提案されており，純粋な LSTM のモデルよりも高い精度が報告されている [96]．Convolutional Neural Network (CNN) [97,98] を用いた手法も提案されている．CNN は画像認識分野で State-of-the-art を記録したモデルであるが，Sentiment Classification 分野でも最高精度を記録している．

これらのモデルの大きな特徴は，特徴量の抽出及び分類モデルの学習を同時に行う点である．従来のモデルよりも高い精度を記録しているため，近年の Sentiment Classification 分野では深層学習を用いた手法が多く提案されいてる．

Distant Supervision に基づく手法

Distant Supervision とは，何らかの手がかりを基に，コストをかけずに大量の擬似ラベル付きデータを収集し，それらを学習させる手法である [99]．Sentiment Classification 分野では特に SNS の投稿の絵文字表現を用いたものが多く行われている．例えば,「今日は晴れた」という tweet に，笑顔を表すような絵文字が付属していた場合，このツイートの極性はポジティブであるとみなすことができる．このようなヒューリスティックによって，あらかじめ各絵文字の極性を定義しておき，その情報を基に自動で各 tweet に擬似的な極性ラベルを割り振る．

擬似ラベル付きのデータセットを構築した後は，そのラベル情報を基にして分類モデルの学習が行われる．その方法は 1.4.2 の教師あり学習に基づく手法で説明したものと同様の手法が用いられる．例えば，Deriu らや Müller らは，CNN を用いた分類モデルを構築している [100, 101]．

1.5 本研究の位置づけ

1.5.1 関連研究との比較

本項では，本研究と目的及び手法が類似している，3 つの分野の既存研究との比較を述べる．

ロボット倫理学分野の研究との比較

1.3 でも述べた通り，ロボット倫理学が扱う対象は多岐にわたる．その中でも特に，本研究はロボットの道徳的知識獲得に焦点を当てたものである．

1.3.2 でも述べた通り，既存の研究は「論理学ベース」及び「ルールベース」のものに大別できる.「論理学ベース」の手法は，道徳を習得させるためのフレームワークを哲学的知見から提案したものである．よって，システムの実装及び精度の検証はされていない.「ルールベース」の手法は，人手により記述された知識やルールを用いた処理が主であり，コストの面から実用的ではない．また，対象となるドメインが限られており，日常生活などの一般的な状況における道徳性に関しては考慮されていない．

上記の問題に対し本研究では，Web 上のテキストデータを用いることで，低コストで幅広い道徳的な常識を獲得することを目的とする．

自然言語処理技術を用いた知識獲得研究との比較

本研究の目的は，Web 上のテキストデータを用いて道徳的な知識を獲得することであり，1.4.1 で述べた知識獲得の研究と目的が類似している．両者の大きな違いは，対象とする知識の種類である．人間と同じように思考するエージェントを作成する上で道徳的な知識は大変重要である．それにも関わらず，既存の知識獲得の研究では，本研究が対象とするような道徳的な知識は対象とされてこなかった．その意味で本研究は，自然言語処理分野における知識獲得研究において，今まで対象とされてこなかった知識を対象とした研究と位置づけることができる．

Sentiment Classification 分野の研究との比較

本研究の入出力は自然言語処理分野の Sentiment Classification における入出力と類似している．具体的には，どちらも自然言語文を入力とし，その文の極性の推定を行う．本研究と既存の Sentiment Classification 分野の研究の大きな違いは，ドメインの差異である．どちらの研究も極性の推定を行うことに変わりはないが，Sentiment Classification では特に製品の評判情報や，人間の感情の推定を行う．それに対し本研究では，人の行動に着目し，その道徳的な善悪性の推定を行う．

1.6 本論文の構成

本論文は全 5 章から成る．

第 2 章では，まず，道徳的な常識の獲得度合いを評価するための道徳判断タスクについて説明すると共に，共起ベースの手法を用いて道徳判断を行う手法を提案する．同時に評価用のデータセットの構築を行い，手法の評価を行う．

第 3 章では，第 2 章で提案した手法の対応言語を英語に拡張した手法の提案を行う．具体的には，分散表現という単語を高次元ベクトルで表現する概念を導入し，それを用いて道徳判断を行う手法を提案する．

第 4 章では，大規模な Web データの解析を行い，道徳的常識を陽に自動獲得する手法について述べる．同時に，獲得した知識を学習することでより高い精度で道徳判断が可能になることを示す．学習の際には，近年機械学習分野で注目されている深層学習が用いられる．特に自然言語処理の様々なタスクで有用であると考えられている注意機構を導入する．また，共起情報を用いることでより高い精度で道徳判断が可能になることを示す．

第 5 章では，本研究において取り組んだ道徳判断タスクについてまとめ，その成果を要約する．また，今後の展望や残された課題についても述べる．

第 2章

評価表現を用いた道徳判断

2.1 序論

推論や判断などの人間が行う高度な情報処理は, 我々が暗黙の内に共有している常識的な知識を元にして行われる [102]. そのため人間の知能を計算機上でシミュレートする際には, 常識の処理は極めて重要な要素であると考えられ, その知識獲得に関する研究が数多くなされている.

Lenat らは, Cyc [103] という常識的知識をデータベース化するプロジェクトを 1984 年に開始した. ここには, CycL と呼ばれる知識記述言語を習得した専門家によって, 常識的知識が手書きで登録されている. Cyc には約 50 万の単語が登録されており, Lenat によると, 成人の持つ常識のうち 2%ほどの常識を蓄えているとされている.

MIT の Open Mind Common Sense(OMCS) [104] では, 一般のボランティアの人々から常識的知識を獲得するアプローチをとっている. OMCS には, 10 年かけて 100 万を超える常識文が登録されている. また, それらを機械による利用可能な形にするため, Concept Net [105] と呼ばれるネットワーク構築を目指している. さらに, 常識的知識の増加を目的として, Analogy Space [106] なる推論機構を導入している. なお, Concept Net 内の概念関係が正しい割合は 6 割強, 推論を加えた場合には 2 割強である.

しかし, これらの常識的知識獲得を目的とした研究には, 次の 3 つの問題点があると考えられる. 人手による知識の登録のため, コストが大きい点 [107], 知識のカテゴリを手動で定めており, 網羅性は未検証である点 [108], 知識が論理形式でのみ記述されており, あいまいな常識を扱っていない点 [108] である. よって, 人手に頼ることなく, あいまいな常識を獲得するシステム構築は大変重要である.

あいまいな常識を獲得することを目的とした研究の中でも, 道徳的常識を人工知能

に実装する研究がある. [109–113]. 例えば, Wallach ら [113] は, 道徳を人工知能に実装するための有益なモデルの作成を目的として研究を行っている. 彼らは, 演繹的モデルと帰納的モデルのハイブリッドモデルを提案している. まず, 演繹的に基本的な指針を定め, 次に細部を帰納的に補強する. しかし, これらの研究 [109–113] は, 提案のみに留まっており, その有効性は示されていない.

一方, Anderson ら [114] は, MedEthEx というシステムを提案している. 彼らは, 医療分野において善悪判断を行うシステムをロボット NAO に組み込んだ. 具体的には, 患者が薬を飲むことを拒否したときに, 患者の意思を尊重するのか, それとも患者の健康を重んじて担当医に連絡をするのかなどの判断を行うシステムである. このシステムでは, まず, 帰納論理プログラミング [115] を用いて, 人間の専門家の推論の事例から帰納的に一般原則を求める. そして次に, その原則をもとに演繹的に善悪の判断を行う. しかし, 専門家による大量の推論事例が必要であるという点及び特定の分野のみにしか適応していないという点には課題が残っていると考えられる.

そこで本章では，Web 上のテキストデータ及び評価表現を用いた道徳判断手法の提案を行う．Web 上には大量のテキストデータが存在しているため，自然言語処理技術を使用することで低コストで道徳常識を獲得できることが期待される．本章における研究の貢献は以下の通りである．

- 道徳判断を行う際に有用である評価表現を自動で獲得する手法を提案する．
- 獲得された評価表現及び Web 上のテキストデータを用いた道徳判断手法を提案する．
- 評価実験を通じ，提案手法の精度検証を行った結果，評価表現選別の有効性を示す．

本章は以下のように構成される．2.2 にて関連研究について概説し，2.3 から 2.6 で評価表現を用いた道徳判断手法について説明する．2.7 で評価実験の結果を述べ，2.8 でまとめを述べる．

2.2 関連研究

2.2.1 共起情報を用いた Sentiment Classification

Sentiment Classification は入力文 (文書) に対し，その極性を出力するタスクである．例えば，商品のレビューを入力として，その商品が好意的に評価されているのか批判的に評価されているのかを出力する．Web 上のテキストデータが増加した 2000 年代前半から研究が行われ始めた．その代表的な手法は，入力文とあらかじめ構築された評価表現辞書を用いるものである [82].

本研究と既存研究の大きな違いは，用いられる評価表現の種類である．既存研究にて構築された評価表現辞書には,「細い」や「魑魅魍魎」といった道徳性には関係がないと考えられる単語が多く含まれており，これらを使用すると道徳判断の精度が大きく下がることが予想される．この問題に対し，道徳判断の際に効果を発揮する評価表現を獲得する手法の提案を行う．

2.2.2 道徳性を備えた人工知能

人工知能やロボットの社会進出の増大に伴い，道徳的に判断し，道徳的に振る舞うことができる人工知能・ロボットの開発が推進されている．この研究は「論理学ベース」のものと「ルールベース」のものに分けることができる.「論理学ベース」の手法は，道徳を習得させるためのフレームワークを哲学的知見から提案したものである [26–28, 109–113]．よって，システムの実装及び精度の検証はされていない.「ルールベース」の手法は，人手により記述された知識やルールを用いた処理が主であり，コストの面から実用的ではない．また，対象となるドメインが限られており，日常生活などの一般的な状況における道徳性に関しては考慮されていない．これらの問題に対し，本研究では Web 上のテキストデータを用いることで，道徳的な常識を低コストで獲得することを目的とする．

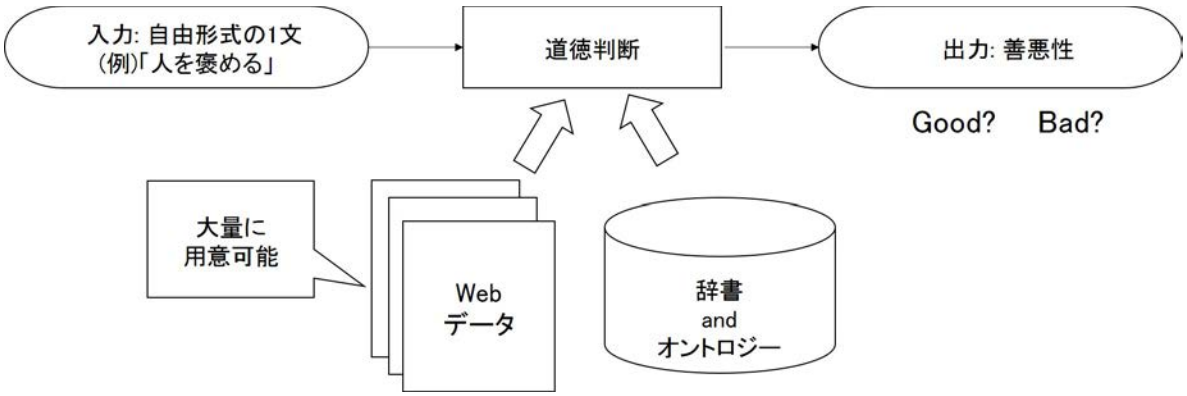

図 2.1 タスクの概要及び本研究のアプローチ

2.3 道徳判断タスク

本研究の目的は，Web 上のテキストデータを用いることで低コストで道徳的な常識を獲得することであるが，その知識の獲得度合いを如何にして評価するかという問題が生じる．この問題に対し，本研究では道徳判断タスクという，自然言語により表現された行為の善悪性を判断するタスクを導入する．

図 2.1 にタスクの概要及び，本研究のアプローチを示す．道徳判断タスクの入力は 1 文であり，その文はある人物の行動を表すものである．出力はその文が表す行為の善悪性である．

道徳判断タスクでは，システムが下した善悪性と，大多数の人間が下した善悪性の一致率を基に評価を行う．これは Wallach らの「機械が持つことのできる能力は機能的道徳性のみであり，それに満足するべきである」という考えに基づいている [113]. また Allen と Wallach らは，機械の道徳性を判定するための「道徳的チューリング・テスト」を提案しており，本タスクも道徳的チューリング・テストの一種であると捉えることができる [116].

2.4 提案手法の概要

提案手法全体の概要を図 2.2 に示す．提案手法は，評価表現抽出フェーズと道徳判断フェーズに分かれる．

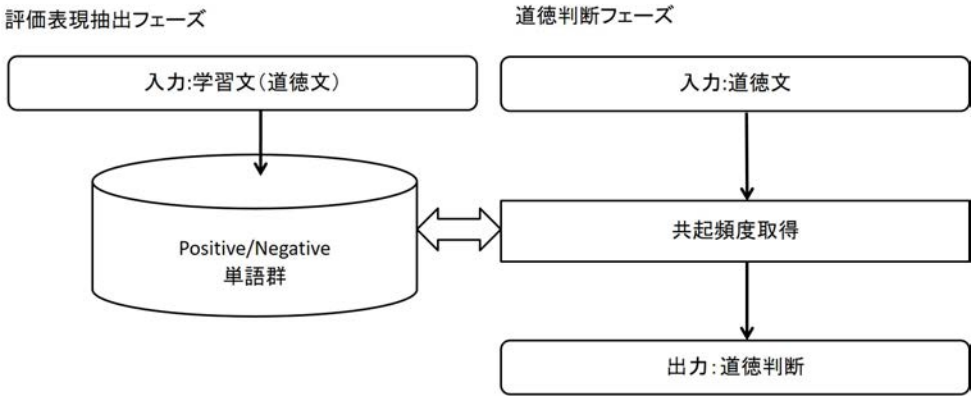

図 2.2 提案手法全体の概要

まず，評価表現抽出フェーズにおいては，道徳判断に適した Positive 単語群及び Negative 単語群を取得することを目的とする．その際，アプレイザル辞書 [77] と評価表現辞書 [72, 73] を組み合わせた辞書（以下，極性辞書とする）の中から Positive 単語群及び Negative 単語群を取得する．具体的には，文部科学省の小学校の道徳の指導要領に書かれている文を基に，それらの文との共起頻度が高い単語が取得される．

次に，道徳判断フェーズの流れを説明する．道徳判断フェーズにおいては，道徳判断が可能な 1 文を入力とする．ここで，道徳判断が可能な文とは,「人を褒める」や「人を殴る」などの善悪判断が可能な文のことである．入力された文と評価表現抽出フェースで獲得した Positive 単語群及び Negative 単語群との共起頻度を比較して道徳判断を行う．なお出力形式は，1.00 から 5.00 の範囲のアナログ値となっており，1.00 に近いほど悪い行為，5.00 に近いほど良い行為となる．

以降，2.5 で評価表現抽出フェーズについて説明し，2.6 で道徳判断フェーズについて説明する．

2.5 評価表現抽出フェーズ

評価表現抽出フェーズでは，道徳判断に適した Positive 単語群及び Negative 単語群を取得することを目的とする．その際，極性辞書の中から Positive 単語群及び Negative 単語群を取得する．文部科学省の小学校の道徳の指導要領に書かれている文，80 文と

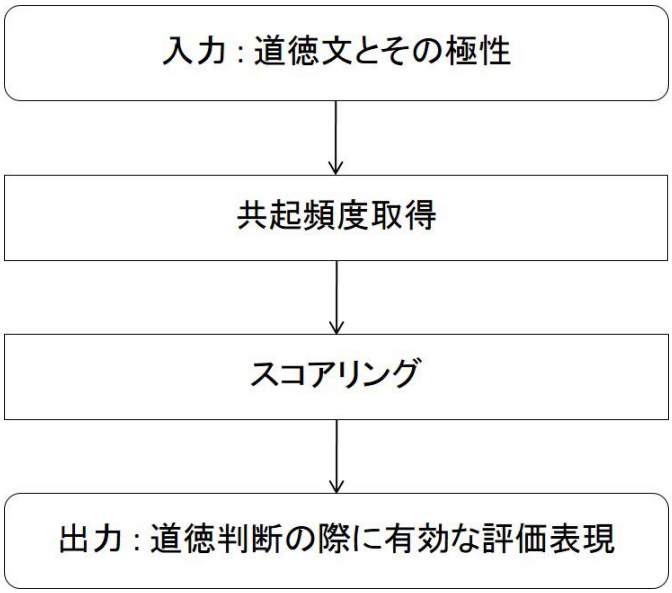

図 2.3 評価表現抽出フェーズの流れ

の共起頻度を基に単語が抽出される．

図 2.3 に評価表現抽出フェーズの流れを示す．

2.5.1 評価表現抽出フェーズが必要な理由

評価表現抽出フェーズが必要な理由は以下の通りである．

(1) 実行時間の短縮

(2) ノイズとなり得る単語群の排除

まず，(1) について説明する．構築した極性辞書には合計 1 万語以上の単語が含まれている．もし道徳判断フェーズにおいて，入力文と全ての単語と共起をとると実行時間の点で現実的ではない．ゆえに，重要な単語群のみを抽出し，実行時間を早める処理が必要になると考えられる．

次に，(2) について説明する．例えば，Positive 単語群には「ため」という語が含まれる．しかし，この単語は「学校を休む “ため” に仮病を使う」というように条件節として使われることもあるため，Negative な文とも共起することが考えられる．そうし

た場合,「ため」という単語はノイズとなり，正確な道徳判断ができなくなることが予想される．ゆえに，そうしたノイズとなりうる単語群を排除する必要性がある．

2.5.2 極性辞書の構築

極性辞書および，アプレイザル辞書を合わせたものを極性辞書として用いる．この辞書の中から道徳判断に適した Positive 単語群及び Negative 単語群を抽出する．

2.5.3 入力文の用意

今回は，文部科学省の小学校の道徳の指導要領に書かれている文を入力として用いる．この指導要領には小学校で学ぶべき道徳が網羅されている．一例として，以下のようなものがある．

- 「健康や安全に気を付け，物や金銭を大切にし，身の回りを整え，わがままをしないで，規則正しい生活をする.」
- 「うそをついたりごまかしたりしないで，素直に伸び伸びと生活する.」

評価表現抽出フェーズにおいてはこれらの文を以下のように 3 文節程度の短い文に区切って用いる．

- 「健康に気をつける」
- 「安全に気をつける」

これは，長い文を入力として用いると，共起頻度を取得するのが難しいと考えられるからである．

また,「嘘をつかない」や「ごまかしたりしない」というような「〜しない」といった否定が入っている文は,「嘘をつく」や「ごまかす」といったように肯定文に直して入力として用いる．

そして,「安全に気をつける」という文章から発展して,「安全運転をする」といったような文章も入力として用いる．

表 2.1 評価表現抽出フェーズにおける入力例

入力文	極性
健康に気をつける	Positive
長所を伸ばす	Positive
節制する	Positive
嘘をつく	Negative
わがままをする	Negative
偏見を持つ	Negative

表 2.2 評価表現抽出フェーズにおける抽出要素例

入力文例	長所を伸ばす
抽出要素例	長所　伸ばす

2.5.4 入力

入力は，2.5.3 で説明した文および，その文の極性である．表 2.1 に入力例を示す．以降,「健康に気をつける」や「嘘をつく」のような文を入力文,「Positive」や「Negative」のことを極性と呼ぶ．今回，極性が Positive である入力文，極性が Negative である入力文をそれぞれ 40 文ずつ用意する．まず，入力文は MeCab [117] により形態素解析され，以下の形態素を抽出要素とする．

(1) 名詞
(2) 動詞
(3) 形容詞

評価表現抽出フェーズにおける抽出要素例を表 2.2 に示す．

表 2.3 評価表現抽出フェーズにおける検索クエリ例

抽出単語	Positve/Negative 単語群
長所 伸ばす	良い
長所 伸ばす	悪い
長所 伸ばす	最高
長所 伸ばす	最低
嘘 つく	良い
嘘 つく	悪い
嘘 つく	最高
嘘 つく	最低

2.5.5 共起頻度取得

次に，抽出された要素と 2.5.2 で構築された極性辞書中の単語が同時に出現しているかを確認する．その際，言語資源としては Web 日本語 N グラムを用いて，7-gram の共起頻度を取得する．検索クエリ例を表 2.3 に示す．

2.5.6 スコアリング

この節では，2.5.5 で取得した共起頻度を元に，極性辞書中の単語にスコアを付ける．スコアリングのアイデアは以下の通りである．

アイデア

(1) 同じ極性の入力文と多く共起する単語に大きなスコアを与える．ただし，異なる極性の入力文と多く共起する単語には大きなペナルティを与える．

(2) 基本的には (1) のスコアを重視する．ただし，同じ順位の単語がある場合，同じ極性の入力文との共起頻度の総和が大きいものに大きなスコアを与える．ただし，

異なる極性の入力文の共起頻度の総和が大きい単語にはペナルティを与える.

まず，(1) について説明する．例えば，Positive な単語「大切」について考える.「大切」という単語は,「約束を守る」や「勇気を持つ」など，極性が Positive である文と共起する．よって，このような単語には大きなスコアを与える．同様に，Positive な単語「ため」も，極性が Positive である文と共起する．しかし，この単語は，極性が Negative である文とも多く共起する．よって，このような単語には大きなペナルティが与えられ，スコアは極めて低くなる.

次に，(2) について説明する．例えば，Positive な単語「尊重」及び「自信」の (1) のスコアが同じだったとする．その場合,「尊重」と Positive な入力文の共起頻度の総和から,「尊重」と Negative な入力文の共起頻度の総和を引いたものと,「自信」についての同様なものを比べ，その値が大きな単語のスコアを高めに設定する.

以下にスコアリングの実際の手順を示す.

手順

(1) 極性辞書中のある Positive な単語 pw_i と共起した極性が Positive である入力文の数を $N_{pw_i,p}$ とする．同様にして，pw_i と共起した極性が Negative である入力文の数を $N_{pw_i,n}$ とする．そして，$ScoreN_{pw_i}$ を以下のように定める.

$$ScoreN_{pw_i} = N_{pw_i,p} - aN_{pw_i,n} \tag{2.1}$$

ここで，a は重みである．これは，Negative な文と共起する Positive 単語に対するペナルティの大きさを示す.

(2) pw_i と極性が Positive である入力文 pl_j との共起数を F_{pw_i,pl_j} とする．同様にして，pw_i 極性が Negative である入力文 nl_k との共起数を F_{pw_i,nl_k} とする．そして，$ScoreF_{pw_i}$ を以下のように定める.

$$ScoreF_{pw_i} = \sum_j F_{pw_i,pl_j} - \sum_k F_{pw_i,nl_k} \tag{2.2}$$

(3) (1) と同様にして，極性辞書中のある Negative な単語 nw_i と共起した Positive 極性の入力文の数を $N_{nw_i,p}$ とする．また，nw_i と共起した Positive 極性の入力文の

表 2.4 評価表現抽出フェーズで取得される Positive 単語例

大切，大事，希望，前向き，信頼，チャンス，新しい，人材， 尊重，自信，恋，きちんと，こよなく，恩恵，風味，重要，安定， 産業，伝統，調和，祖先，喜び，知恵，爽やか，完全，支え

数を $N_{nw_i,p}$ とする．そして，nw_i のスコア $ScoreN_{nw_i}$ を以下のように定める．

$$ScoreN_{nw_i} = N_{nw_i,n} - aN_{nw_i,p} \tag{2.3}$$

ここで (1) と同様に，a は重みである．これは，Positive な文と共起する Negative 単語に対するペナルティの大きさを示す．

(4) (2) と同様にして，nw_i と極性が Positive である入力文 pl_j との共起数を F_{nw_i,pl_j} とする．同様にして，nw_i 極性が Negative である入力文 nl_k との共起数を F_{nw_i,nl_k} とする．そして，$ScoreF_{pw_i}$ を以下のように定める．

$$ScoreF_{nw_i} = \sum_k F_{nw_i,nl_k} - \sum_j F_{nw_i,pl_j} \tag{2.4}$$

例えば，Positive 辞書に登録されている「必要」という単語について考える．この単語は，極性が Positive である文「整理整頓をする」と共起する．しかし，この単語は，極性が Negative である文「無駄遣いをする」とも共起する．よってこの場合,「必要」という単語は，道徳判断においてノイズとなる可能性があると考えられる．そういった単語へのペナルティが式 (2.1) や式 (2.3) における $aN_{pw_i,n}$ や $aN_{nw_i,p}$ の項である．

2.5.7 出力: 重要な Positive 単語群/Negative 単語群

2.5.6 で算出された $ScoreN_{pw_i}$，$ScoreN_{nw_i}$ が大きい順に単語が取得される．ただし，$ScoreN_{pw_i}$，$ScoreN_{nw_i}$ の大きさが同じ単語に関しては，$ScoreF_{pw_i}$，$ScoreF_{nw_i}$ が大きいものを優先して取得する．抽出される Positive 単語例，Negative 単語例は表 2.4，表 2.5 のようなものがある．

表 2.5 評価表現抽出フェーズで取得される Negative 単語例

ダメ，嫌い，勝手，馬鹿，最低，禁止，嘘つき，嫌，嫌う， 大嫌い，下手，暴力，危ない，罪悪，病気，デマ，イヤ，苦手， 神経質，出任せ，悲しい，恐ろしい，痛い，心苦しい，ショック

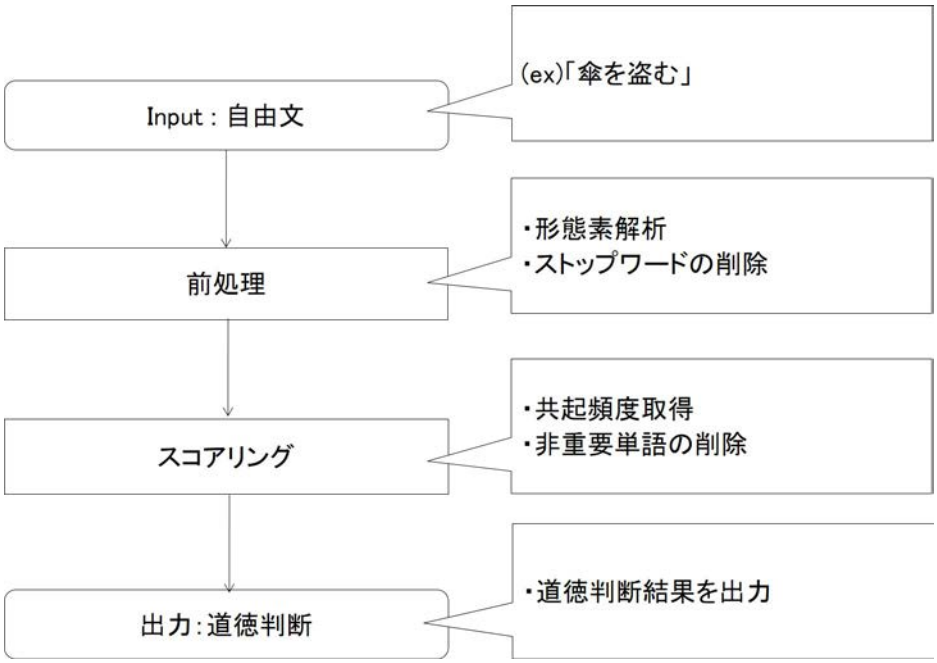

図 2.4 道徳判断フェーズの流れ

2.6 道徳判断フェーズ

道徳判断フェーズでは 2.5 で抽出した Positive 単語群及び Negative 単語群を元に，ユーザが入力した文の道徳判断をすることを目的とする．

図 2.4 に道徳判断フェーズの流れを示す．

2.6.1 入力

道徳判断フェーズにおける入力は自由形式の 1 文である．

表 2.6 ストップワード

する，ある，よる，いる，なる，いう，みる，できる

表 2.7 道徳判断フェーズにおける抽出要素例

入力文例	電車内で電話する
抽出要素例	電車　電話

2.6.2 前処理

入力文は MeCab [117] により形態素解析され，以下の形態素が抽出される．

(1) 名詞（一般，固有名詞，サ変接続，形容動詞語幹）

(2) 動詞（自立）

(3) 形容詞（自立）

ただし，表 2.6 に示す 8 形態素は出現頻度が非常に多く，また他の語句の補助的な機能を持つ語句であることから，ストップワードとする [118]．ここで，ストップワードとは，抽出候補形態素にその語句が入っていても，抽出要素としない語句を集めたものである．ここで，文中に含まれている否定語「ない」の個数を d として数えておく．この d は最終的なスコアリングの際に用いられる．

道徳判断フェーズにおける抽出要素例を表 2.7 に示す．

2.6.3 共起頻度取得

2.6.2 で抽出された要素と 2.5 で抽出された Positive 単語群及び Negative 単語群の共起頻度を取得する．その際，言語資源としては Web 日本語 N グラムを用い，7-gram での共起頻度を取得する．

ここで，共起頻度が取得できれば 2.6.5 で説明するスコアリングに進む．もしも共起頻度が取得できなければ，2.6.4 で説明する入力文の非重要単語の削除を行う．

2.6.4 非重要単語の削除

非重要単語の削除は 2.6.3 節で，共起頻度が取得できなかった場合に行われる．ここでの目的は，2.6.2 節で抽出した要素の中で，重要度の低い要素を省くことである．

以下に，理想的な処理例を示す．

入力例 「自分 行動 責任 持つ」

処理例 「行動 責任 持つ」

非重要単語の削除は以下の手順で行われる．なお，基本的なアルゴリズムは，文書中に含まれる単語の重要度を評価する tf-idf 法を参考にし，以下の 2 点を考慮している．

- 1 文中に同じ単語が複数出てくれば，その単語に大きな重みを与える．
- 言語資源中に頻出する単語であれば，その単語には小さな重みを与える．

手順

(1) 入力された単語群の中で，重複している単語があれば 1 つにまとめ，重複している数を n とする．このとき，単語が重複していなければ $n = 1$ となる．

(2) 単語群の中で，i 番目の単語 w_i が Web 日本語 N グラム中に出現する頻度を F_{w_i} とする．

(3) 各単語毎の重み W_{w_i} を式 (2.5) によって算出する．

$$W_{w_i} = \frac{n}{F_{w_i}} \tag{2.5}$$

(4) W_{w_i} が一番小さい要素を取り除く．

非重要単語の削除が終わったら，また 2.3.4 節の共起頻度取得に進む．以降，共起頻度が取得できるまで，非重要単語の削除を繰り返す．ただし，要素が 1 つになるまで非重要単語の削除が行われる．

2.6.5 スコアリング

取得した共起頻度を元に，以下のアイデアに基づいてスコアリングが行われる．

アイデア

(1) 入力文と極性が Positive である単語との共起頻度の総和を求める．ただし，共起頻度は，その単語が言語資源中に出現する頻度で正規化する．

(2) (1) と同様にして，入力文と極性が Negative である単語との共起頻度の総和を求める．ただし，共起頻度は，その単語が言語資源中に出現する頻度で正規化する．

(3) (1) で求められる値から (2) で与えられる値を引く．

実際のスコアリングの手順は以下の通りである．

手順

(1) 抽出要素と，ある Positive 単語 pw_i との共起頻度を F_{pw_i} とし，pw_i が言語資源中に現れる頻度を $F_{pw_{all}}$ とする．そして，式 (2.6) によって，抽出要素と Positive 単語との共起頻度を正規化した後，足し合わせる．

$$PScore = \frac{1}{F_{pw_{all}}} \sum_{i=1} F_{pw_i} \tag{2.6}$$

(2) 同様に，抽出要素と，ある Negative 単語 nw_i との共起頻度を F_{nw_i} とし，nw_i が言語資源中に現れる頻度を $F_{nw_{all}}$ とする．そして，式 (2.7) によって，抽出要素と Negative 単語との共起頻度を正規化した後，足し合わせる．

$$NScore = \frac{1}{F_{nw_{all}}} \sum_{i=1} F_{nw_i} \tag{2.7}$$

(3) 以下の式によって最終的なスコアを決定する．

$$Score = \begin{cases} \frac{PScore - NSscore}{PScore + NSscore} \times 2.00 + 3.00 & \text{(否定単語の数 } d \text{ が偶数のとき)} \\ -\frac{PScore - NSscore}{PScore + NSscore} \times 2.00 + 3.00 & \text{(否定単語の数 } d \text{ が奇数のとき)} \end{cases} \tag{2.8}$$

ここで，分母の $PScore + NSscore$ および 2.00, 3.00 は正規化のために用いている．この結果，$Score$ は 1.00 から 5.00 の範囲の値をとる．

2.6.6 出力

出力は，式 (2.8) の $Score$ となる．$Score$ は，1.00 から 5.00 の範囲の実数となっており，1.00 に近いほど悪い行為，5.00 に近いほど良い行為となる．

2.7 評価実験

2.7.1 実験目的

評価表現を用いた道徳判断システムの有効性の検証を行う．

2.7.2 実験概要

システムが出力した道徳判断が，人間の道徳判断の結果と整合性が取れているかどうかを評価する．その際入力文はあらかじめ，8 人から 100 文程度を集めておく．そして，入力文に対する道徳判断を以下の 5 段階で人手による評価を行う．

1 とても悪い
2 どちらかと言えば悪い
3 どちらか判断がつかない
4 どちらかと言えば良い
5 とても良い

2.7.3 実験条件

8 名の被験者それぞれから 10 文〜30 文程度の道徳文を収集した．収集された道徳文は合計 150 文あったが，その中には重複している文や，表現がおかしい文，道徳判断として相応しくない文などが存在した．よって，それらの文を省いた．その結果，117

文が残り，それについて 5 段階での評価を 23 人 (20 代男性 12 人，20 代女性 11 人) の被験者に行ってもらった．なお評価の際は，117 文をランダムで表示した．この際，評価表現抽出フェーズにおける重み a は 2 とした．また，システムの出力は 1.00 から 5.00 に正規化される．

2.7.4 実験結果

表 2.8 に入出力例を示す．表 2.9 に 2 値判断における再現率・適合率・F 値を示す．ただし，人間による評価値の平均が 3.00 以上の事例を Positive な事例，3.00 未満の事例を Negative な事例とした．また，システムの出力も同様に扱った．適合率，再現率，F 値は以下の式によって求められる．

$$\text{適合率} = \frac{tp}{tp + fp} \tag{2.9}$$

$$\text{再現率} = \frac{tp}{tp + fn} \tag{2.10}$$

$$F\text{ 値} = \frac{2 \times \text{適合率} \times \text{再現率}}{\text{適合率} + \text{再現率}} \tag{2.11}$$

ここで，tp, fp, tn, fn は以下の事例数を表す.

- tp (true positive): システムが positive と判断して, 正解データも positive であった事例数.
- fp (false positive): システムが positive と判断したが，正解データでは positive ではなかった事例数.
- tn (true negative): システムが negative と判断して，正解データも negative であった事例数.
- fn (false negative): システムが negative と判断したが，正解データでは positive であった事例数.

表 2.8 入出力例

入力文	システムの出力	人間の評価の平均
自転車を盗む	1.00	1.00
救急車に道を譲る	5.00	5.00
障害者をいたわる	4.47	4.45
仏壇を蹴る	1.17	1.26
お年寄りの荷物を持つ	4.62	4.65
墓参りをする	4.71	4.61
信号無視をする	1.36	1.48
傘を盗む	1.58	1.13
他人の携帯を見る	1.00	1.78
ごみを分別する	4.25	4.87
公共の建物内で静かにする	5.00	4.65
植物に水をやる	3.91	4.56
食べながら図書館の本を読む	5.00	1.61
人から貰ったプレゼントを他人にあげる	5.00	1.91

図 2.5 に人間の評価の平均とシステムの道徳判断の関係を示す．人間の評価の平均とシステムの出力の相関係数 0.65 であった．図 2.6 に人間の評価の平均とシステムの道徳判断の関係を示す．ただし，評価表現抽出を行わず，あらかじめ構築されている評価表現辞書全ての単語との共起頻度を計算し，道徳判断を行った結果である．この場合，人間の評価の平均とシステムの出力の相関係数 0.34 であった．

表 2.10 に評価表現抽出フェーズを用いた場合と用いなかった場合の相関係数と二値判断における F 値を示す．

表 2.9 再現率・適合率・F 値 (2 値)

単語削除回数	再現率	適合率	F 値
0	0.31(36/117)	0.90(36/40)	0.46
0-1	0.62(72/117)	0.86(72/84)	0.72
0-2	0.77(90/117)	0.79(90/114)	0.78
無制限	0.80(93/117)	0.80(93/117)	0.80

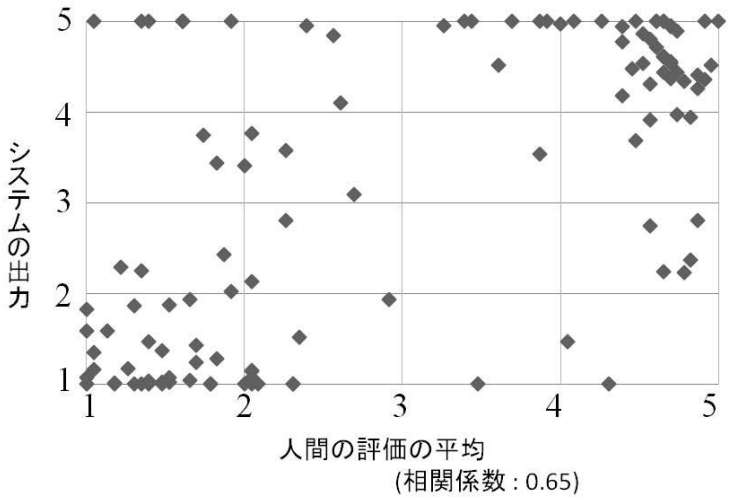

図 2.5 システムと人間の道徳判断の相関係数

2.7.5 考察

表 2.10 から，評価表現を選別した場合は，しない場合に比べて精度が向上していることが分かる．このことから，評価表現抽出フェーズにより，道徳判断の際に重要な表現を抽出出来たことが示唆される．

図 2.7 から，単語の削除回数が増加すると共に，道徳判断の精度が低下することが分かる．これは，単語を削除する回数が多ければ多いほど，元の文の意味を逸脱してしまうためであると考えられる．

図 2.5 から不正解の事例は左上に固まっていることが分かる．このことから，人間が Negative だと思っている事例を，システムは Positive と出力しやすいということが

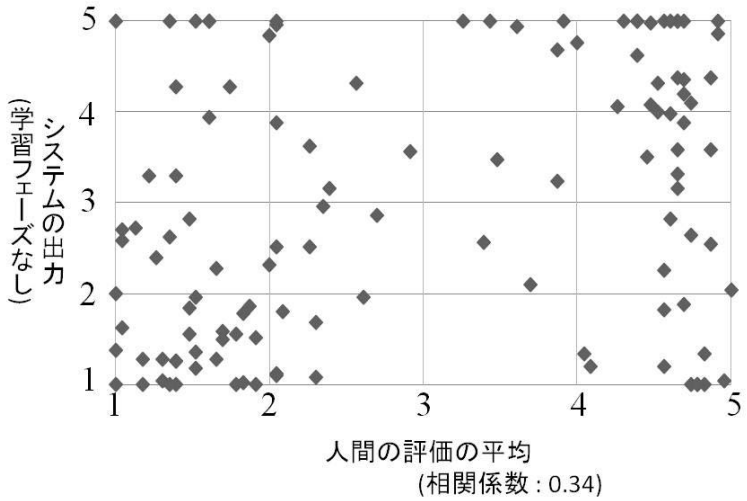

図 2.6 システムと人間の道徳判断の相関係数

表 2.10 評価表現抽出を用いた場合と用いなかった場合の比較

評価表現抽出	相関係数	F 値 (二値判断)
○	0.65	0.80
×	0.34	0.69

分かる．例えば,「食べながら図書館で本を読む.」や「人から貰ったプレゼントを他人にあげる」などの入力文に対しては適切な判断ができなかった．これは，提案手法が単語と単語の繋がりを考慮せず，Bag-of-Words として入力文を扱っていることが原因であると考えられる．具体的には,「食べながら図書館で本を読む.」という文からから抽出された単語群は，Positive な評価表現と共起しやすい．それゆえ，システムが誤った判断をしたと考えられる．

2.8 評価表現を用いた道徳判断システムのまとめ

本章では，評価表現を用いた道徳判断システムを提案した．基本的には共起頻度を基にしたシステムであり，あらかじめ Positive な単語群及び Negative な単語群を定義

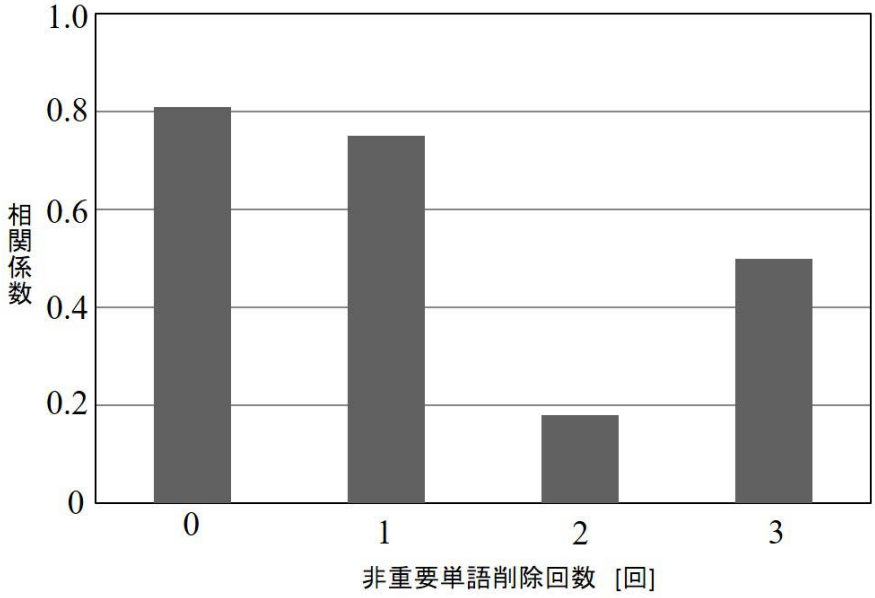

図 2.7 非重要単語削除回数と相関係数の関係

し，それらと入力文の共起情報を基に道徳的な善悪性の判断を行う．ここで，あらかじめ構築されている評価表現辞書にはノイズとなるような単語が含まれる．そこで，評価表現を抽出する手法を提案し，道徳判断の際に効果を発揮する評価表現を獲得を行った．

第 3章

分散表現を用いた道徳判断

3.1 序論

本章では，英語を対象言語とした際の道徳判断手法について考える．第 2 章で提案した手法及び，使用したデータは日本語を対象言語としたものであるため，その手法をそのまま用いることは困難である．特に，共起頻度を計算する際に利用した Web 日本語 N-gram は隣接 7-gram の共起頻度を計算できたが，英語版の N-gram コーパスでは隣接 5-gram の共起頻度のみしか計算できない．それに加え，第 2 章で提案した手法は評価表現を用いて道徳判断を行うものであるが，入力文として単語数が少ない文章を想定しており，単語数が多い文章に対しては道徳判断が困難である．

そこで本章では，分散表現の利用，ならびに連想情報の利用により上記の問題点の改善を図る．ここで，分散表現とは実数値のベクトル表現のことであり，本章における研究では入力文の分散表現化を行う．分散表現を用いて自然言語文を表現することで，文全体の考慮や文と文の類似度の計算などが容易になる．具体的にはまず，Mikolov らの手法 [119] により単語の分散表現が学習される．その後評価表現辞書を用いて，道徳判断タスクに特化した単語の分散表現が学習される．文が入力された際には，学習された単語の分散表現を用いて入力文の分散表現が得られる．

言葉の背景にある常識の考慮に関しては，道徳判断の際，入力文から連想される情報を利用する．人間はテキストに直接的に明示されていない情報を連想して文章を理解していると考えられる．そのような連想情報を道徳判断システムに組み込むことで，より精度の高い道徳判断が可能になることが期待される．具体的には，まず構文解析器 Enju [120] を用いて，Britich National Corpus (BNC) [121] から述語項構造が抽出される．次に，述語項構造の分散表現が計算され，述語項構造と分散表現から成るデー

タベースが作成される．文が入力された際には，入力文との類似度が高い述語項構造がデータベース中から抽出される．その後，入力文の情報と共に，抽出された連想による述語項構造を用いた道徳判断が行われる．

本章における研究の貢献は以下の通りである．

- 道徳判断タスク向けの単語の分散表現の再学習手法を提案する．
- 再学習された単語の分散表現および連想情報を用いた道徳判断システムを提案する．
- 評価実験により，特に長い入力文に対する道徳判断の精度が向上することを示す．

以降，3.2 で関連研究について述べ，3.3 で道徳判断タスクに向けた単語の再学習手法について説明する．その後，3.4 で提案手法について詳説し，3.5 で評価実験の結果を，3.6 でまとめを述べる．

3.2 関連研究

3.2.1 単語の分散表現

Mikolov らによって提案された単語の分散表現の学習手法は，その利便性から大きな注目を集めた．この手法によって学習された分散表現は，似た意味を持つ単語同士の分散表現が似るだけではなく，“vector(king) - vector(man) + vector(woman) = vector(queen)” のような意味の演算が可能であるという報告がされている．この手法は word2vec としてツール化されており，2018 年現在でも様々なタスクに用いられている．なお，単語の分散表現の学習方法については付録 B に示す．

word2vec を改良したモデルも数多く提案されている．Pennington らは，Glove という word2vec よりも大域的な情報を考慮できる分散表現手法を提案した [122]．Facebook からは 2016 年に FastText という未知語に対応できるモデルが提案されている [123]．理論的な解析も進んでおり，幾つかの前提を置くと word2vec のモデルが shifted PMI 行列の分解と等価であることが示されている [124]．

一方，単語の分散表現には数多くの課題が存在する．一般に論じられている問題には，語義性の問題がある．例えば，“bank” という単語には “土手” と “銀行” という 2 つの意味があるが，“bank” には単一の分散表現を割り当てることが多い．この問題に対して，Neelakantan らは単語毎の語義数を自動で決定しその分散表現を学習する手法を提案している [125].

他には，極性の問題が広く知られている．word2vec の学習の特性上，“good” と “bad” は似たような分散表現が割り当てられる可能性が高い．それゆえ，Sentiment Classification などの極性推定タスクでは精度を下げる要因になり得る [126]．本研究で扱う道徳判断タスクも極性推定のタスクの 1 つであるため，word2vec により学習された分散表現をそのまま利用すると精度が下がる要因になると考えられる．この問題に対して本研究では，単語の分散表現を道徳判断タスク向けに再学習する手法の提案を行う．これにより，分散表現を用いて道徳判断を行うことが可能となる．具体的な手法については，3.3 で詳しく説明する．

3.3 道徳判断タスクに特化した分散表現の学習

3.3.1 概要

まず，Mikolov らの手法 [119] を用いて，平文コーパスから単語の分散表現が学習される．彼らの手法では，注目している単語の周辺語が予測できるようにニューラルネットワークが学習され，各単語が低次元のベクトルで表現される．この手法により学習された単語の分散表現は，自然言語処理の様々なタスクで有用であることが示されている [127–129].

しかしながら，文献 [119] で学習されるのは汎用の分散表現であり，タスクに特化したものではない．よって，その分散表現をそのまま特定のタスクに用いてしまうと精度低下の原因になることがある [126]．例えば，単語 “good” と単語 “bad” の周辺文脈は似ていることが多く，文献 [119] で学習された上記 2 単語の分散表現の類似度は高くなってしまう．従って，文献 [119] の手法により得られた分散表現を本研究のような善悪判断にそのまま用いると精度が低下してしまう可能性がある．

表 3.1 辞書に収録されている単語例.

Positive 単語例
honesty, kindness, justice, devotion, ethical, modesty, optimism, loyalty, patience, confidence, sincere, respect, appreciation, friendship, purity, fairness, importance, fairness
Negative 単語例
unethical, dangerous, careless, inconsiderate, incompetent, Immoral, hate, rude, dishonest, irrational, improper, foolish, Ignorant, irrelevant, jerk, unsafe, unhappy, shameful, war

そこで，Mikolov らの手法で学習された単語の分散表現と 2 章と似たような手法で構築された辞書を用いて，道徳判断タスクに特化した単語の分散表現の学習を行なう．以降，この分散表現の再学習手法について詳説する．

3.3.2 学習データ

学習データとして，2 章と同様の手法で構築された英語版の辞書を用いる．辞書の中には 300 単語 (Positive: 150 単語，Negative: 150 単語) が収録されており，これらを再学習用のデータとして用いる．辞書に収録されている単語例を表 3.1 に示す．

3.3.3 単語の分散表現の変換

前述した単語辞書を基に，道徳判断タスクに特化した分散表現の学習を行なう．具体的には，以下の非線形変換により，既存の単語の分散表現が道徳判断タスク用の分散表現に変換される．

$$\boldsymbol{x} = \alpha \cdot \tanh(\frac{1}{\alpha} \cdot \boldsymbol{M} \cdot \boldsymbol{x'}) \tag{3.1}$$

ここで，α はスケーリング用の定数であり，$\boldsymbol{x}'$，$\boldsymbol{x}$ はそれぞれ元の単語の分散表現，変換後の分散表現を表す．$\boldsymbol{M}$ は単語の分散表現を変換するための行列である．また，関数 tanh は要素毎に施される．

行列 $\boldsymbol{M}$ を学習するために，学習データの分類問題をロジスティック回帰で表現する．変換後の単語ベクトル $\boldsymbol{x}$ が Positive になる確率は以下の通りになる．

$$P(y = +1 \mid \boldsymbol{x}) = \frac{1}{1 + \exp(-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x})} \tag{3.2}$$

ここで，$\boldsymbol{w}$ は入力ベクトル と内積が計算されるパラメータベクトルであり，学習対象である．T はベクトルの転置を表す．また，式 (3.2) を用いて，入力単語ベクトル $\boldsymbol{x}$ が Negative になる確率は以下の式で表される．

$$P(y = -1 \mid \boldsymbol{x}) = 1 - P(y = +1 \mid \boldsymbol{x}) = 1 - \frac{1}{1 + \exp(-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x})} \tag{3.3}$$

目的関数を考える際には式が 1 つのほうが扱いやすいので，式 (3.2) と式 (3.3) をまとめて以下のように表す．

$$P(y \mid \boldsymbol{x}) = \frac{1}{1 + \exp(-y\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x})} \tag{3.4}$$

ここで，y は各単語のラベルであり，単語の極性が Positive の場合 $+1$ となり，Negative の場合 -1 となる．学習では，次の目的関数 $L(\boldsymbol{w}, \boldsymbol{M})$ を最小にするパラメータ $\boldsymbol{w}$, $\boldsymbol{M}$ を求める．

$$L(\boldsymbol{w}, \boldsymbol{M}) = -\sum_{i=1}^{N} \log p(y^{(i)} \mid \boldsymbol{x}^{(i)}) + c_1 r_1(\boldsymbol{w}) + c_2 r_2(\boldsymbol{M}) \tag{3.5}$$

ここで c_1, c_2 はハイパーパラメータである．また，N は学習データ数である．$r_1(\boldsymbol{w})$ は通常の L2 正則化であり，以下の式で表される．

$$r_1(\boldsymbol{w}) = \|\boldsymbol{w}\|^2 \tag{3.6}$$

$r_2(\boldsymbol{M})$ は行列 $\boldsymbol{M}$ を単位行列 に近くなるように正則化する関数であり，以下の式で表される．

$$r_2(\boldsymbol{M}) = \|\boldsymbol{M} - \boldsymbol{I}\|^2 \tag{3.7}$$

ここで，$\|...\|$ はフロベニウスノルムを表す．

3.3.4 パラメータの学習

式 (3.1) の tanh の影響で，式 (3.5) の最適化問題は非凸である．そこで，確率的勾配降下法により局所最適解を求める．具体的には，$t+1$ 反復時点でのパラメータ $\boldsymbol{w}^{(t+1)}$, $\boldsymbol{M}^{(t+1)}$ は，t 反復時点でのパラメータを用いて以下の式で求められる．

$$\boldsymbol{w}^{(t+1)} = \boldsymbol{w}^{(t)} - \eta \frac{\partial L(\boldsymbol{w}, \boldsymbol{M})}{\partial \boldsymbol{w}} \tag{3.8}$$

$$\boldsymbol{M}^{(t+1)} = \boldsymbol{M}^{(t)} - \eta \frac{\partial L(\boldsymbol{w}, \boldsymbol{M})}{\partial \boldsymbol{M}} \tag{3.9}$$

ここで，η は学習率である．

3.4 分散表現を用いた道徳判断手法

図 3.1 に道徳判断システムのフローチャートを示す．入力は自由文形式の英語の文章 1 文である．文が入力されたら，まず 3.3 で学習された単語の分散表現を用いて，入力文の分散表現が計算される．分散表現を用いて自然言語文を表現することで，文と文の類似度の推定などが容易になり，単語数が多い文章に対応可能になることが期待される．

次にあらかじめ構築された述語項構造データベースの中から後に説明するような連想情報が取得される．最後に，入力文情報および連想情報を用いて道徳判断が行なわれる．連想情報をシステムに組み込むことで精度の高い道徳判断が可能になることが期待される．出力は 1.00 から 5.00 の範囲のアナログ値であり，1.00 に近いほど悪い事例，5.00 に近いほど善い事例であると判断される．

以降，連想情報を取得する際に使用する述語項構造データベースの構築方法について 3.4.1 で述べ，3.4.2 で道徳判断システムについて述べる．

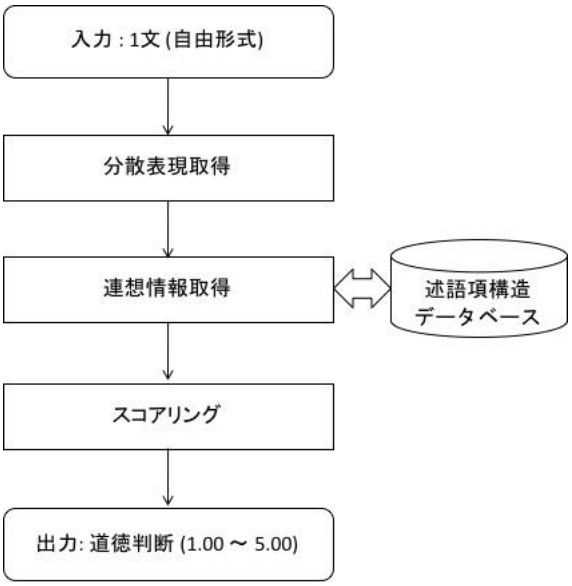

図 3.1 道徳判断の流れ

3.4.1 述語項データベースの構築

本節では，連想の際に用いる述語項構造データベースの構築方法について述べる．まず，英文構文解析器 Enju [120] により，平文コーパスが構文解析される．Enju の出力結果には述語項構造解析結果が存在しており，これを用いて述語項構造データベースの構築が行なわれる．提案システムでは，Enju の述語項構造解析結果の中でも特に，以下の述語項構造の抽出が行なわれる．

- 1 項をとる動詞で，項が名詞句か代名詞
- 2 項をとる動詞で，項が名詞句か代名詞
- 3 項をとる動詞で，項が名詞句か代名詞

なお，単語は原形で抽出される．例えば，Enju を用いて “He ran the department and bought a candy.” を構文解析すると，表 3.2 のような結果が得られる．また，述語項構造の構成要素の単語ベクトルの和がその述語項構造のベクトルとなる．つまり，述語項構造中に存在する単語の分散表現を $\boldsymbol{b}_i$ とすると，述語項構造の分散表現 $\boldsymbol{p}$ は以下の式から求められる．

$$\boldsymbol{p} = \sum_{i=1}^{N_p} \boldsymbol{b}_i \tag{3.10}$$

表 3.2 Enju による述語項構造解析結果の例.

述語	項 1	項 2
run	he	department
buy	he	candy

ここで，N_p は述語項構造中に含まれる単語の数である．また，各単語の分散表現は 3.3 で学習された分散表現が用いられる．

3.4.2 道徳判断システム

入力

自由文形式の英語の文章 1 文を入力とする．文が入力されると，Stanford Core NLP [130] を用いて，入力文の原形化が行なわれる．このとき，入力文に含まれる否定単語の数を d とする．ここで，否定単語とは，“not” や “never” などの文章を否定形にする働きを持つ単語のことである．この否定単語の数 d は，3.4.2 のスコアリングの際に用いられる．

分散表現取得

その後，入力文の分散表現が計算される．入力文中に含まれる，否定単語を除いた単語の分散表現を $\boldsymbol{e}_i$ とすると，入力文の分散表現 $\boldsymbol{s}$ は以下の式から求められる．

$$\boldsymbol{s} = \sum_{i=1}^{N_s} \boldsymbol{e}_i \tag{3.11}$$

ここで，N_s は入力文中に含まれる単語の数である．また，各単語の分散表現は 3.3 で学習された分散表現が用いられる．

連想情報取得

次に，入力文の分散表現を基に連想情報の取得が行なわれる．まず，3.4.1 で構築された述語項構造データベース中の全ての要素に対し，入力文とのコサイン類似度が計算される．入力文の分散表現を $\boldsymbol{s}$，述語項構造の分散表現を $\boldsymbol{p}$ とすると，コサイン類似度は以下の式により求められる．

$$cos(\boldsymbol{s}, \boldsymbol{p}) = \frac{\boldsymbol{s} \cdot \boldsymbol{p}}{|\boldsymbol{s}| \cdot |\boldsymbol{p}|} \tag{3.12}$$

$$= \frac{\sum_{i=1}^{N_d} s_i p_i}{\sqrt{\sum_{i=1}^{N_d} s_i^2} \cdot \sqrt{\sum_{i=1}^{N_d} p_i^2}} \tag{3.13}$$

ここで，N_d は分散表現の次元数である．式 (3.12) により計算されるコサイン類似度が閾値 δ 以上の述語項構造が連想情報として取得される．

スコアリング

連想情報の取得が行われた後，入力文の情報と連想情報を用いて道徳判断が行なわれる．入力文が Positive である確率 P_{prob} は以下の式で求められる．

$$P_{prob} = \frac{\sigma(\boldsymbol{w}^{\mathrm{T}} \boldsymbol{s}) + \sigma(\boldsymbol{w}^{\mathrm{T}} \boldsymbol{a})}{2} \tag{3.14}$$

ここで，$\boldsymbol{w}$ は 3.3 で学習されたパラメータである．また，$\boldsymbol{a}$ は，3.4.2 で取得された連想情報の分散表現の平均であり，以下の式により求められる．

$$\boldsymbol{a} = \frac{\sum_{i=1}^{N_a} \boldsymbol{p}_i}{N_a} \tag{3.15}$$

ここで，N_a は入力文から連想情報として取得された述語項構造の数である．$\sigma(x)$ はシグモイド関数であり，以下で定義される．

$$\sigma(x) = \frac{1}{1 + \exp(-x)} \tag{3.16}$$

また，P_{prob} は 0.00 から 1.00 の範囲の確率なので，入力文が Negative である確率 N_{prob} は以下の式により計算される．

$$N_{prob} = 1 - P_{prob} \tag{3.17}$$

最終的なスコアは以下の式によって求められる．

$$Score = \begin{cases} (P_{prob} - N_{prob}) \times 2.00 + 3.00 & (\text{否定単語の数 } d \text{ が偶数のとき}) \\ -(P_{prob} - N_{prob}) \times 2.00 + 3.00 & (\text{否定単語の数 } d \text{ が奇数のとき}) \end{cases} \tag{3.18}$$

出力

出力は式 (3.18) の $Score$ となる．$Score$ は 1.00 から 5.00 の範囲の実数となっており，1.00 に近いほど悪い行為，5.00 に近いほど善い行為となる．

3.5 評価実験

3.5.1 評価用データセット

評価実験には以下の 2 つのデータセットを用いた．

TestSet1 1 つ目のデータセットは，第 2 章で作成されたデータセットを英訳したものである．全 115 文から成り，各文について主観評価実験により得られた 1.00 から 5.00 のアナログ値で正解が付与されているデータセット (平均単語数 5.27) である．

TestSet2 TestSet1 に加え，単語数が多い文章から成るデータセット (平均単語数 8.27) を作成し実験を行った．具体的な作成方法は以下の通りである．

- 日本語 Web コーパス 2010 [131] の一部 (2GB) から言語パタンを用いて道徳に関係があると考えられる文章をランダムに 500 文 (Positive: 250 文，Negative: 250 文) 収集した．まず，道徳的に Positive と考えられる文章の収集に関しては,「ことは良い」という言語パタンに係る動詞およびその周辺文脈が収集される．道徳的に Negative と考えられる文章の収集に関しては,「ことは悪い」という言語パタンに係る動詞およびその周辺文脈が収集される．図 2 の例であれば,「人を褒める」が Positive な候補として,「悪口を言う」が Negative な候補として収集される．

- 収集された文章を英訳し，単語数が 6 以上の文章のみを抽出する．

- 抽出された文章を用いて 20 代の被験者 12 名に「文として意味をなさないと考えられる文章」および「道徳判断が困難であると考えられる文章」を削除してもらった．
- 上記の結果，合計 47 文が残った．それらの文について同様の被験者に以下の基準で善悪判断を行ってもらった．

  1. 悪い
  2. どちらかといえば悪い
  3. どちらともいえない
  4. どちらかといえば善い
  5. 善い

3.5.2 実験設定

分散表現初期値

道徳判断タスクに特化した分散表現を学習する際の分散表現初期値として，1,000 億単語からなるニュース記事から skip-gram モデルを用いて学習された単語の分散表現を使用した．このデータは Mikolov らにより公開されており，各単語は 300 次元のベクトルで表現されている [132]．

パラメータ・学習回数

表 3.3 に評価実験で使用したパラメータを示す．学習の際，パラメータベクトル $\boldsymbol{w}$ は，0.00 から 1.00 の範囲の乱数を初期値とし，行列 $\boldsymbol{M}$ は単位行列を初期値とした．また，学習回数は 1,000 回とした．

使用する平文コーパス

述語項構造を取得する際には，British National Corpus (BNC) [121] を用いた．BNC は，書き言葉，話し言葉合わせて 1 億語から成る，イギリス英語コーパスである．ま

表 3.3 評価実験の際に使用したパラメータ.

パラメータ	値
スケーリング用パラメータ α	1.00
$r_1(\boldsymbol{w})$ のスケーリング用パラメータ c_1	1/600
$r_1(\boldsymbol{w})$ のスケーリング用パラメータ c_1	1/600
学習率	0.01
閾値 δ	0.60

ず，BNC の元のファイルから，一行一文として，6,020,399 行を取得することができた．それらの文に対して，3.4.2 で説明した構文解析器 Enju を用いて構文解析を行った結果，合計 35,859 件の述語項構造が取得された．

比較手法

以下の手法を比較手法として用いる．

- Co-occurrence: 第 2 章で提案した手法を英語版に拡張したものである．
- ベースライン: 提案手法において，単語の分散表現を再学習せず，また連想情報も使用しない場合をベースラインの手法とする．学習はロジスティック回帰により行なわれる．つまり，3.3.2 の学習データを用いて，以下の目的関数を最小化するようなパラメータ $\boldsymbol{w}$ を学習する．

$$L(\boldsymbol{w}) = -\sum_{i=1}^{N} \log p(y^{(i)} \mid \boldsymbol{x}^{(i)}) + c_1 r_1(\boldsymbol{w}) \tag{3.19}$$

そして，式 (3.19) の最小化により学習された $\boldsymbol{w}$ を用いて，入力文が Positive である確率を以下の式により求める．

$$P_{prob} = \sigma(\boldsymbol{w}^{\mathrm{T}} \boldsymbol{s}) \tag{3.20}$$

ここで，$\boldsymbol{s}$ は式 3.11 から得られる入力文の分散表現である．入力文が Negative である確率は式 (3.17) により計算される．最終的なスコアは，式 (3.18) により求

表 3.4 手法毎の相関係数

手法	相関係数	
	TestSet1	TestSet2
Co-occurrence	0.43	0.11
ベースライン	0.49	0.40
Re-embedding	0.53	0.47
提案手法	0.54	0.52

められる．

- Re-embedding: 提案手法において，連想情報を使用しない場合の手法である．つまり，入力文が Positive である確率は式 (3.20) により求められる．また，入力文が Negative である確率は式 (3.17) により計算される．最終的なスコアは，式 (3.18) により求められる．

3.5.3 実験結果

表 3.4 に各手法で得られた結果と，正解データとの相関係数を示す．TestSet1 においては，正解データと提案手法の出力の相関係数は 0.54 であった．TestSet2 においては，正解データと提案手法の相関係数は 0.52 であった．図 3.2 に TestSet1 における正解データと提案手法により得られた出力の関係を示す．また，図 3.3 に TestSet2 における正解データと提案手法により得られた出力の関係を示す．

表 3.5 に TestSet1 における二値判断の適合率，再現率，F 値を示す．提案手法は正解率 0.78 で 2 値判断が可能であった．また，表 3.6 に TestSet2 における二値判断の適合率，再現率，F 値を示す．提案手法は正解率 0.77 で 2 値判断が可能であった．ここで，正解率は以下の式によって求められる．

$$\text{正解率} = \frac{tp + tn}{tp + fp + tn + fn} \tag{3.21}$$

ここで，tp, fp, tn, fn は第 2 章と同様に定義される．

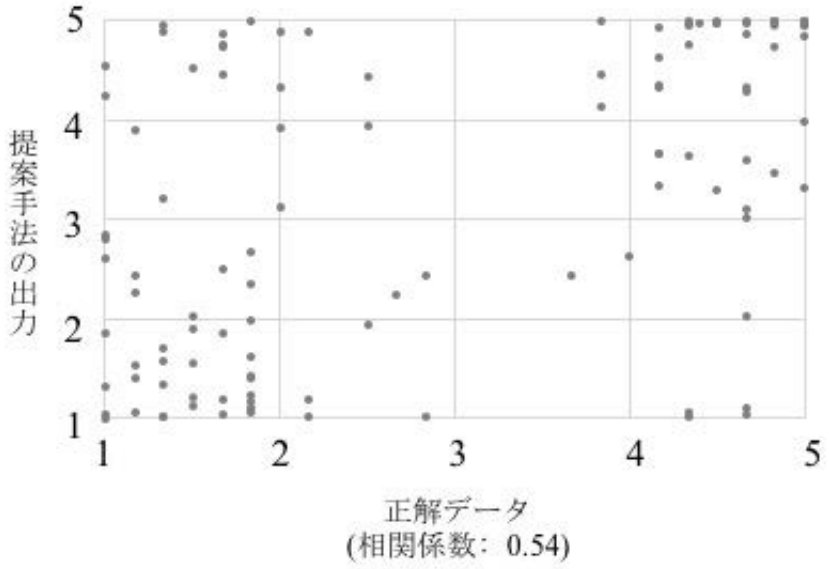

図 3.2 正解データとシステムの出力の関係 (TestSet1)

3.5.4 考察

分散表現の再学習の影響

TestSet1　表 3.4 より，正解データと単語の分散表現を再学習しない場合 (ベースライン) の相関係数は 0.49 である．それに対して，単語の分散表現を再学習した場合 (Re-embedding) の相関係数は 0.53 であった．また，表 3.5 よりベースラインの場合，正解率 0.67 で二値判断が可能であり，Re-embedding の場合，正解率 0.72 で二値判断が可能であることが分かる．以上より，単語の分散表現を道徳判断タスク用に再学習することにより，道徳判断の精度が上がったと考えられる．

TestSet2 表 3.4 より，正解データと単語の分散表現を再学習しない場合 (ベースライン) の相関係数は 0.40 である．それに対して，単語の分散表現を再学習した場合 (Re-embedding) の相関係数は 0.47 であった．また，表 3.6 よりベースラインの場合，正解率 0.66 で二値判断が可能であり，Re-embedding の場合，正解率 0.71 で二値判断が可能であることが分かる．以上より，TestSet2 においても，単語の分散表現を道徳判断タスク用に再学習することにより，道徳判断の精度が上がったと考えられる．

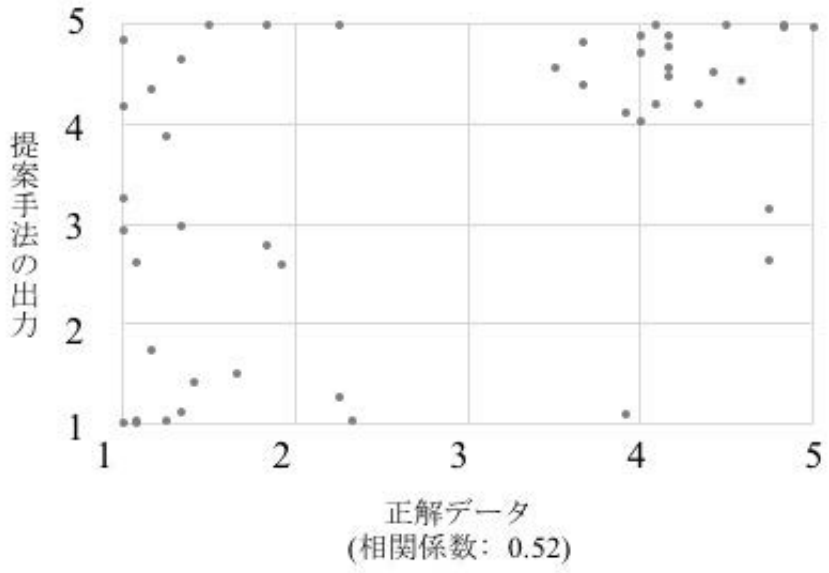

図 3.3 正解データとシステムの出力の関係 (TestSet2)

連想情報の利用の影響

TestSet1 表 3.4 より，正解データと提案手法の相関係数は 0.54 である．表 3.5 より，提案手法は正解率 0.78 で二値判断が可能である．それに対し，上記で述べたように，Re-embedding の場合，相関係数は 0.53，二値判断の正解率は 0.72 である．以上より，連想情報の利用により，道徳判断の精度が上がったと考えられる．

TestSet2 表 3.4 より，正解データと提案手法の相関係数は 0.52 である．表 3.6 より，提案手法は正解率 0.77 で二値判断が可能である．それに対し，上記で述べたように，Re-embedding の場合，相関係数は 0.47，二値判断の正解率は 0.71 である．以上より，連想情報の利用により，道徳判断の精度が上がったと考えられる．

Co-occurrence の手法と提案手法の比較

TestSet1 表 3.4 より正解データと Co-occurrence の手法の相関係数は 0.43 である．また，表 3.5 より，Co-occurrence の手法は正解率 0.73 で二値判断が可能である．それに対し，提案手法の場合，相関係数は 0.54，二値判断の正解率は 0.78 である．相関係

表 3.5 二値判断の正解率・適合率・再現率・F 値 (Testset1)

	正解率	適合率	再現率	F 値
Co-occurrence	0.73 (84/115)	0.70 (39/56)	0.74 (39/53)	0.72
ベースライン	0.67 (77/115)	0.59 (47/79)	0.89 (47/53)	0.71
Re-embedding	0.72 (83/115)	0.66 (43/65)	0.81 (43/53)	0.73
提案手法	0.78 (90/115)	0.70 (45/64)	0.85 (45/53)	0.77

数に関しては 0.11 の向上が見られたものの，二値判断の正解率に関しては大幅な向上は見られなかった．このことから，単語数が少ない文章に対しての道徳判断の精度は提案手法，Co-occurrence の手法で大きく変わらないことが示唆される．

TestSet2 表 3.4 より正解データと Co-occurrence の手法の相関係数は 0.11 である．また，表 3.6 より，Co-occurrence の手法は正解率 0.57 で二値判断が可能である．それに対し，上記で述べたように，提案手法の場合，相関係数は 0.52，二値判断の正解率は 0.77 である．相関係数に関しては，0.41 の向上が見られ，二値判断に関する正解率に関しても 0.20 の向上が確認された．このことから，単語数が多い文章に対しての道徳判断の精度は，Co-occurrence の手法に比べ，提案手法のほうが優れていることが示唆される．

3.5.5 エラー分析

提案手法において，誤った道徳判断が行われた事例についての分析を行った．図 3.2 および図 3.3 より，図の右下の部分には点が少ないが，左上の部分には点が多いことが分かる．このことから以下の 2 点が示唆される．

- 「善い事例」を「悪い事例」と判断することは少ない．
- 「悪い事例」を「善い事例」と判断することは多い．

表 3.6 二値判断の正解率・適合率・再現率・F 値 (Testset2)

	正解率	適合率	再現率	F 値
Co-occurrence	0.57 (27/47)	0.56 (14/25)	0.61 (14/23)	0.58
ベースライン	0.66 (31/47)	0.61 (20/33)	0.87 (20/23)	0.71
Reembedding	0.71 (33/47)	0.65 (20/31)	0.87 (20/23)	0.74
提案手法	0.77 (36/47)	0.71 (21/30)	0.91 (21/23)	0.79

実際に,「善い事例」を「悪い事例」と判断したのは，10 事例なのに対し,「悪い事例」を「善い事例」と判断したのは，26 事例だった．そこで，その 26 事例に着目してエラー分析を行った．表 3.7 に提案手法が誤って善い事例と判断した文の例を示す．このように大規模な常識および高度な推論が必要な事例に対し，誤った道徳判断が行なわれることが多かった．例えば,“I read a library’s book while eating. (食べながら図書館の本を読む.)”という文の善悪判断を行なうためには,「食べ物をこぼすかもしれない.」という常識だけではなく,「食べ物をこぼしたら本が汚れるかもしれない.」という推論が必要となる．また,「図書館の本は公共の物である.」という常識や「公共の物を汚すことは善くない.」という常識も必要となると考えられる．上記のような常識と推論を組み合わせることで，この事例の善悪判断が可能となると考えられる．しかしながら，現状では提案手法において上記程に高度な推論を行なうことは困難であると考えられる．それ故，そのような事例に対して善悪判断を誤ってしまったと考えられる．この問題に対しては，知識獲得機構および連想機構をさらに強化することで対処することが考えられる．

表 3.7 提案手法が誤って善い事例と判断した文の例.

入力文	提案手法の出力
I read a library's book while eating. (食べながら図書館の本を読む.)	4.32
I cut in a line. (列に割り込む.)	3.22
I talk while driving. (運転しながら喋る.)	3.94
I kick the statue of Buddha. (仏像を蹴る.)	4.53
I give a gift from my friend to others. (友達から貰ったプレゼントを, 他の人に渡す.)	4.94
I use a cellular phone on the priority seat. (優先席で携帯を使用する.)	4.89

3.6 分散表現を用いた道徳判断システムのまとめ

本章では，分散表現を用いた道徳判断手法を提案した．提案手法の大きな特長は文の分散表現化である．これにより単語数が多い文章に対しても道徳判断を行うことが可能となった．また分散表現化の際，道徳判断タスクに特化させることにより，システム全体の性能の向上を図った．評価実験では，人間による道徳判断と提案手法による道徳判断の比較が行われた．その結果，提案手法の有効性が示唆された．

第 4章

自動獲得された擬似ラベル付きデータを用いた道徳判断

4.1 序論

Sentiment Classification とは，与えられた入力文 (文書) に対し，その極性を推定するタスクである．現在までに様々な手法が提案されてきているが，多くの手法は学習データを用いない手法と，学習データを用いる手法に大別することができる．学習データを用いない手法では，共起情報を用いて極性の推定を行う．Turney らは，対象とする文と “excellent” や “poor” などの評価表現との共起情報を用いる Sentiment Classification を行う手法を提案した [82]．学習データを用いる場合は，機械学習により入力と出力の関係を学習する．Pang らは，評判分析を初めて機械学習の問題として定式化した [88]．その後，様々な研究が行われているが，最近では深層学習による手法が盛んに研究されている [94, 96, 97]．

一般的には，対象ドメインの学習データを大量に用意し，それらを基に機械学習モデルを構築することが精度向上に繋がると考えられている．よって，道徳判断タスクでも道徳性に関係のあるドメインのデータを集め，機械学習モデルを構築することで精度が向上することが期待される．しかしながら，道徳的な事例についてラベル付けされたデータは存在しない．また，人手でのラベル付けはコストの問題に加え，知識の網羅性の問題も生じるため現実的ではない．

そこで本章では，擬似ラベル付きデータを自動獲得し，それらのデータを用いた道徳判断手法を提案する．提案手法は大きく 2 つに分けることができる．まず，擬似ラベル付きデータの自動獲得である．ここでは，ある文に対して Positive もしくは Negative

のラベルがつけられたデータを自動で獲得することを目的とする．具体的には，評価表現，接続表現，構文情報などの言語的パタンを用いて，これらの擬似ラベル付きデータが収集される．

次に獲得されたデータを用いて機械学習モデルが構築される．具体的には，N-gram 特徴量を基にしたロジスティック回帰モデルおよび，注意機構を導入した深層学習によるモデルである．獲得されたデータはある文章に対してラベルが付与されているため，それらを訓練データとして分類モデルの構築が行われる．

本章における研究の貢献は以下の通りである．

- 道徳的な知識を Web 上のテキストデータから自動で獲得する手法を提案する．
- 獲得された知識を訓練データとして用いて道徳判断を行う手法を提案する．
- 評価実験を通して，提案手法による精度向上を確認する．

以降，4.2 で関連研究について説明し，4.4 で擬似ラベル付きデータを自動獲得する手法について述べる．4.5 で自動獲得された擬似ラベル付きデータを用いた道徳判断手法について説明し，4.6 及び 4.7 で評価実験の結果を，4.8 でまとめを述べる．

4.2 関連研究

4.2.1 擬似ラベル付きデータを自動獲得する研究

自然言語を入力とし，それがどのようなカテゴリに属するかを判断するモデルを機械学習により構築する研究は数多くなされている．最も代表的なタスクは入力された文書がどのカテゴリに属するかを予測する文書分類タスクであり，本研究と類似している Sentiment Classification タスクもこの分類に属する．一般的に文書分類タスクにおいては，あらかじめ人手で構築されたラベル付きデータを基に入力文書とラベル間の対応関係を学習する．しかしながら，大量の文書にラベル付けを行うことはコストが高い．そこで，人手によってラベル付けしたデータではなく，ルールなどを用いて半自動でラベル付けされたデータを獲得し，それらを用いて機械学習モデルの構築を行う研究がある．

知識ベース拡張のための関係抽出というタスクでは，テキストを入力としてそこからトリプルと呼ばれるエンティティ間の関係を獲得することを目標とする．例えば，以下のような入出力を考える．

- **入力**: Obama was born in United States, who worked as a USA president.
- **出力**: born-in (Obama, USA)

このようなある文に対して関係ラベルが付与されたデータが大量にあることが理想的であるが，ラベル付きデータを大量に作成することはコストの面から困難である．そのため，近年ではDistant Supervisionと呼ばれる手法が盛んに研究されている [42–46]．これは「知識ベース中の既知の関係を示す2つのエンティティを含む文は，その関係を説明する」という仮定に基づき，2つのエンティティを含む文に対して擬似的な関係のラベルを付与し，それを訓練データとして用いるというものである．

その他，本研究と似たようなタスク設定の研究も行われている．鍛冶らは，HTML文書からルールやパタンなどを用いて極性タグ付きコーパスを自動で獲得し，評判分析を行うモデルを提案している [79]．徳久らは，Webから感情生起要因コーパスを構築し，そのデータを基に感情推定を行うモデルの提案を行っている [133]．

これらの研究と本研究の違いは，ドメインの差異である．既存研究では感情や評判などを対象としており，道徳性について扱った研究は存在しない．そのため，既存研究の手法をただ適応するだけでは道徳性に関係のないデータまで収集されてしまう．具体的には，以下の2つの問題が生じる．

- あらかじめ構築されている評価表現辞書を用いると，道徳性に関係のないデータが獲得されてしまう点．
- 人の行動を表していないデータが獲得されてしまい，データの中にノイズが生じてしまう点．

1つ目の問題に対しては第2章で構築された評価表現辞書を用いることで対応する．2つ目の問題に対しては，入力文が道徳判断可能な文か否かの判定を行う機構を提案し，ノイズとなるようなデータを削除することで対応する．

4.2.2 深層学習

深層学習とは，多層のニューラルネットワークによる機械学習手法であり，近年様々な分野で過去の精度を上回る報告がなされている．

機械翻訳の分野では，2014 年に提案された Encoder-Decoder モデルが従来の Phrase Based 機械翻訳とほぼ同等の性能であると報告されている [134]．従来の手法では，様々な処理を統合して翻訳を行っていたのに対し，Encoder-Decoder モデルでは，End-to-End で学習が可能であることも注目すべき大きな特長である．2015 年には注意機構の発見によって，既存手法の精度を大きく上回る報告がされている [13,14]．2016 年には Google の翻訳システムも深層学習を用いたものに置き換わり，その高品質な翻訳の実現が世間に大きな衝撃を与えた [15]．

機械翻訳分野だけではなく，自然言語処理のあらゆる分野で深層学習は用いられている．応用タスクでは，テキスト自動要約 [135]，評判分析 [95]，質問応答 [136]，音声認識 [137]，基礎的な解析タスクでは，品詞タグ付け [138] や構文解析 [139] などの幅広い分野で深層学習が用いられている．

社会的にも大きく注目されており，数々の技術が実用化されている．2015 年には Microsoft 社が「りんな」というチャット AI の開発を行った [140]．サービス開始から 1 か月でユーザ数が 130 万人を超えるなど，そのクオリティの高いキャラクタ性，対話返答能力などが注目を集めた．医療分野では，IBM の Watson が患者の病状を見抜くだけではなく，その治療方法を適切に割り出した事例が報告されている．その他，Watson には画像認識 [141]，音声認識 [142]，性格分析 [143] など様々な機能が存在する．

深層学習の登場以降，マルチモーダルな情報処理を必要とする研究・実用化事例も増加している．入力された画像からその説明文を出力する Image Caption の研究は，言語情報処理と画像情報処理を組み合わせた代表的な研究例である [144,145]．他にも，ロボットの情報処理や自動運転技術，マルチモーダルな対話システムなどの研究が行われている．

本章では，多様な分野で有効とされている深層学習が道徳判断タスクにおいても有効であるかの検証を行う．それと同時に，共起情報も組み合わせることでより有意な特徴量を抽出することが可能であるかを検証する．

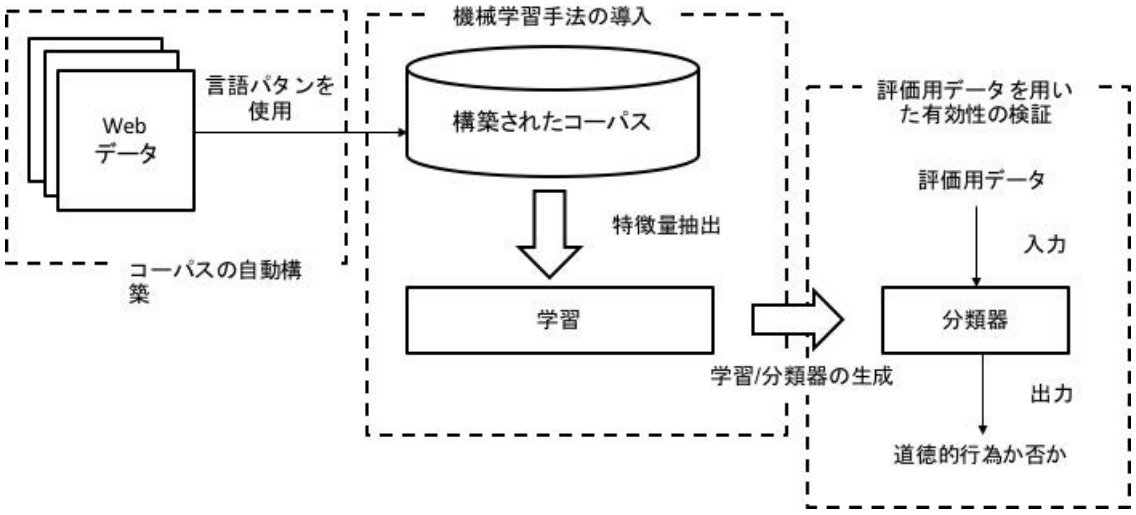

図 4.1 本章で提案する手法の全体図.

4.2.3 深層学習による Sentiment Classification

Sentiment Classification 分野でも深層学習によって精度が向上することが多数報告されている．第 1 章で説明した通り，様々なモデルが提案されてきているが，Recursive Neural Network [91, 92] を用いた研究や Recursive Tensor Network [93] を用いた研究などが存在する．

本研究と最も類似した研究は Wang らの研究である [95]．彼らは時系列ニューラルネットワークに注意機構を導入したモデルを用いて極性の推定を行った．本研究でもこのモデルと類似したモデルを用いて道徳判断を行う．本研究と Wang らの研究の違いは，以下の 2 点である．1 つ目は，人手によってラベル付けされたデータを用いるのではなく，擬似的にラベルが付与されたデータを用いる点である．2 つ目は，先行研究のモデルを用いるだけではなく，本タスクで重要であると考えられる共起情報を導入し，その精度の検証を行う点である．

4.3 本章で提案するシステムの概要

図 4.3 に本章で提案する手法の全体図を示す．提案手法ではまず，擬似ラベル付きデータの自動獲得が行われる．次に，獲得されたデータを基に機械学習モデルの構築

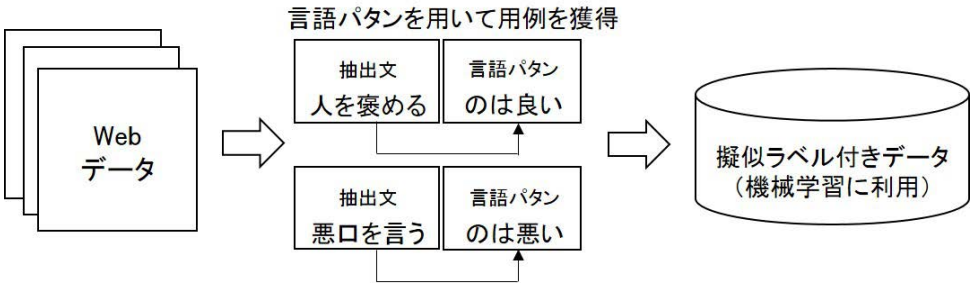

図 4.2 擬似ラベル付きデータ自動獲得の基本的なアイデア.

が行われる．最後に，学習された機械学習モデルを基に道徳判断が行われ，その性能が評価される.

4.4 擬似ラベル付きデータの自動獲得

本節では，擬似ラベル付きデータを自動獲得する手法について説明する．本手法の基本的なアイデアを図 4.2 に示す．基本的には，評価表現及び接続表現，構文情報を用いてデータの獲得を行う．例えば，図 4.2 の例では「人を褒める」行為が Positive な事例,「悪口を言う」行為が Negative な事例として獲得できる.

図 4.3 に擬似ラベル付きデータを自動獲得する際のフローチャートを示す．具体的には，言語パタンや評価表現を用いて，獲得候補が抽出される．これらの獲得候補の中にはノイズとなるような事例が多数存在するため，それらを除去する処理が行われる．最終的にはある文に対して “Positive” もしくは “Negative” のラベルが付与されたデータが獲得される.

4.4.1 入力

擬似ラベル付きデータを獲得する際の入力は，用意されたコーパス中に収録されている文全てである.

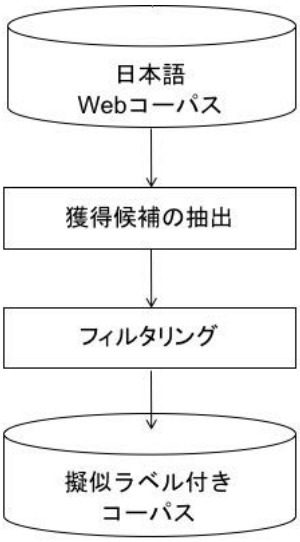

図 4.3 擬似ラベル付きデータを自動獲得する際のフローチャート.

4.4.2 獲得候補の抽出

入力された文毎に以下の処理が行われる.

1. 構文解析器 CaboCha [146] を用いて構文解析が行われる.
2. あらかじめ定義された接続表現が評価表現に係っているかの判定が行われる. もし係っていなかった場合, これ以降の処理は省略され, 次の文に対して解析が行われる.
3. 接続表現が評価表現に係っていた場合, 接続表現に係っている形態素が抽出される.
4. 抽出された形態素に係っている形態素が抽出される.
5. 抽出された全ての形態素を元の文の順番に並べ, 獲得候補とする. また極性は, 評価表現に定義されている極性を参照する.

この処理の際に用いられる接続表現および評価表現の例を表 4.1 に示す. 評価表現には 2 章で構築された辞書に収録されているものが用いられる. また, 接続表現には “ことは” や “のは” など 4 種類のパタンが用いられる.

表 4.1 接続表現及び評価表現の例.

	個数	具体例
Positive な評価表現	54	良い，尊い
Negative な評価表現	56	悪い，酷い
接続表現	4	ことは，のは

表 4.2 意志動詞の例.

愛する，信じる，楽しむ，食べる，使う，行く，待つ

4.4.3 フィルタリング

獲得された候補の中には道徳判断の際にノイズとなるようなものが存在する．本節では，そのようなノイズとなるような事例を除去する手法について説明する．図 4.4 にノイズ除去手法のフローチャートを示す．ノイズの除去は以下の手順で行われる．

1. 獲得候補の一番最後の形態素の原型が意志動詞であるかの判断が行われる．ここで，意志動詞とは人間の意志による動作を表す動詞である．意志動詞の例を表 4.2 に示す．意志動詞ではなかった場合，ノイズとして獲得候補から削除される．

2. もし意志動詞であった場合，文の主語が人を表す単語，もしくは何もないかの判定が行われる．この条件に当てはまらない場合，ノイズとして獲得候補から削除される．ここで，人を表す単語はあらかじめ定義されているものが用いられる．人を表す単語の例を表 4.3 に示す．

表 4.3 人に関する単語の例.

私，僕，自分，俺，我，彼，あいつ，奴

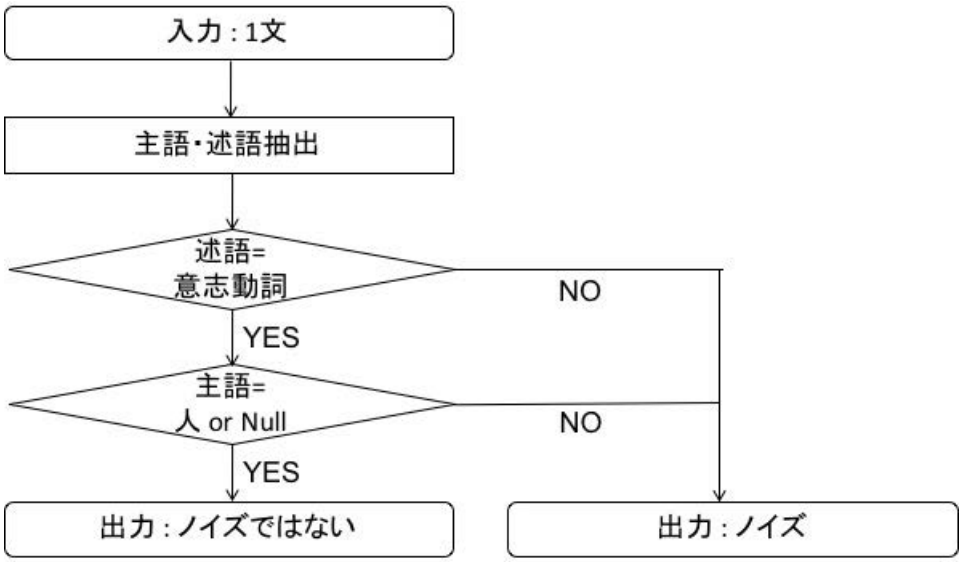

図 4.4 ノイズ除去手法のフローチャート.

4.4.4 出力

出力は擬似ラベル付きデータの集合である．具体的には，各文に対して “Positive” もしくは “Negative” のどちらかのラベルが付与された事例の集合が出力となる.

4.5 擬似ラベル付きデータを用いた道徳判断手法

本節では，4.4 において自動獲得された擬似ラベル付きデータを用いた道徳判断手法について説明する．アイデアとしては，前節で用いたデータを訓練データとみなし，教師ありの機械学習を行う.

図 4.5 に本章で提案するネットワークを示す．提案ネットワークは以下の 2 つのネットワークに分けることができる．1 つ目は分散表現学習ネットワーク (4.5.2) である．このネットワークを用いて，入力文から特徴量の抽出が行われる．2 つ目は道徳判断ネットワーク (4.5.3) である．このネットワークを用いて，入力文の特徴量及び共起情報を考慮した道徳判断が行われる.

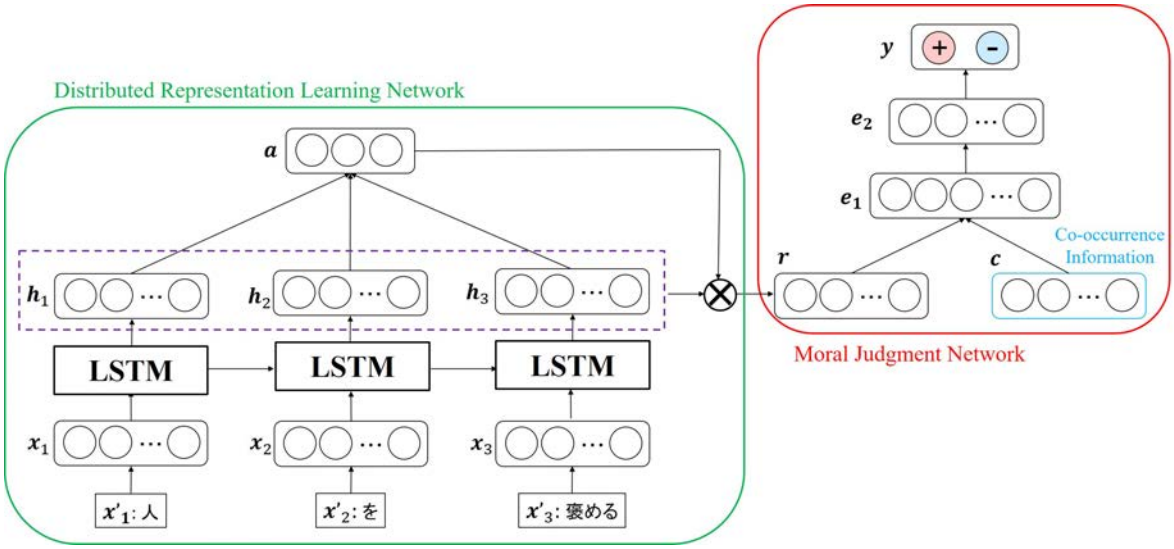

図 4.5 本章で提案するネットワーク.

4.5.1 入力

入力は前章までと同様に 1 文である．入力文は，形態素解析器 MeCab [117] によって形態素単位に分かち書きが行われる．

4.5.2 分散表現学習ネットワーク

分かち書きされた形態素はその順番のままネットワークの入力となる．

$$\boldsymbol{x}_t = \boldsymbol{W}_{x'x}\boldsymbol{x'}_t. \tag{4.1}$$

ここで，$\boldsymbol{x'}_t$ は時刻 t に入力された形態素に対応する次元のみが 1 で，残りは 0 を取る 1-hot vector である．また $\boldsymbol{x}_t$ は第一層の出力であり，$\boldsymbol{W}_{x'x}$ はパラメータ行列である．

第一層目の出力 $\boldsymbol{x}_t$ は第二層目の Recurrent Neural Network(RNN) への入力となる．ここで，RNN の各セルには Long Short-Term Memory(LSTM) セルが用いられる．LSTM は RNN の勾配消失や勾配爆発の問題を防ぐために拡張されたものである．

図 4.6 に LSTM のセルを示す．LSTM の特徴は 3 つのゲート関数と 1 つのメモリセルである．ゲートには入力ゲート，忘却ゲート，出力ゲートが存在し，入力された情

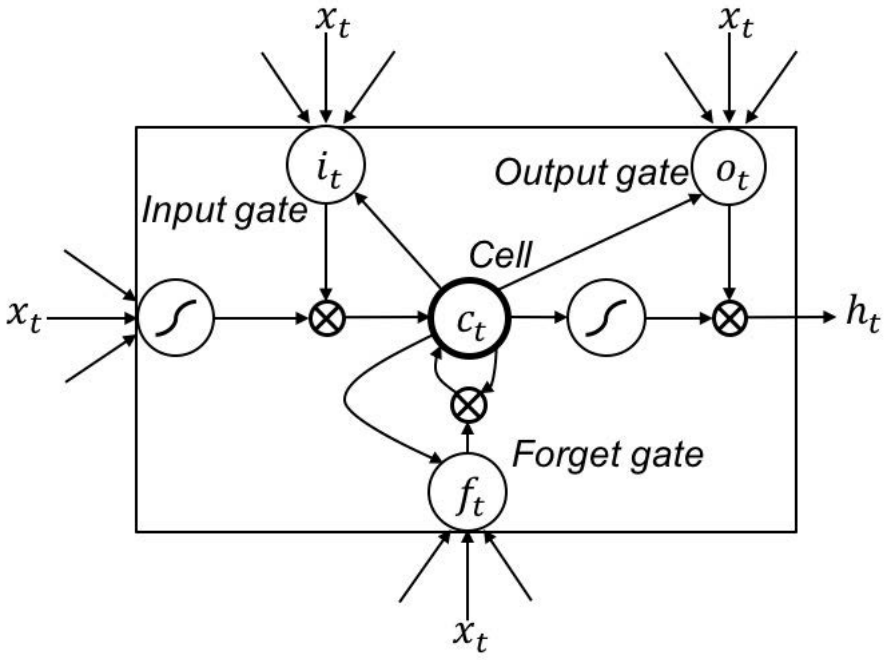

図 4.6 Long Short-Term Memory Cell [3].

報に対し各ゲートが作用することで，上述した勾配消失や勾配爆発の問題を解決する．具体的には，以下の通りに定式化される．

$$\boldsymbol{i}_t = \sigma(\boldsymbol{W}_{xi}\boldsymbol{x}_t + \boldsymbol{W}_{hi}\boldsymbol{h}_{t-1} + \boldsymbol{b}_i) \tag{4.2}$$

$$\boldsymbol{f}_t = \sigma(\boldsymbol{W}_{xf}\boldsymbol{x}_t + \boldsymbol{W}_{hf}\boldsymbol{h}_{t-1} + \boldsymbol{b}_f) \tag{4.3}$$

$$\boldsymbol{c}_t = \boldsymbol{f}_t\boldsymbol{c}_{t-1}\boldsymbol{i}_t \tanh(\boldsymbol{W}_{xc}\boldsymbol{x}_t + \boldsymbol{W}_{hc}\boldsymbol{h}_{t-1} + \boldsymbol{b}_c) \tag{4.4}$$

$$\boldsymbol{o}_t = \sigma(\boldsymbol{W}_{xo}\boldsymbol{x}_t + \boldsymbol{W}_{ho}\boldsymbol{h}_{t-1} + \boldsymbol{W}_{co}\boldsymbol{c}_t + \boldsymbol{b}_o) \tag{4.5}$$

$$\boldsymbol{h}_t = \boldsymbol{o}_t \tanh(\boldsymbol{c}_t) \tag{4.6}$$

ここで，σ は各要素に作用するシグモイド関数であり，tanh はハイパボリックタンジェントを表す．$\boldsymbol{W}_{**}$ 及び $\boldsymbol{b}_*$ は LSTM のパラメータである．最終的な出力は $\boldsymbol{h}_t$ で定義される．

注意機構のパラメータ a_t 及び，式 (4.6) で計算される $\boldsymbol{h}_t$ を用いて最終的な出力は以

下の通りになる．

$$\boldsymbol{r} = \sum_{t=1}^{N} a_t \boldsymbol{h}_t \tag{4.7}$$

ここで，注意機構のパラメータ a_t は順伝搬ネットワークにより計算される．

$$d_t = \boldsymbol{v} \cdot \tanh(\boldsymbol{W}_{h_t} \boldsymbol{h}_t + \boldsymbol{b}_d) \tag{4.8}$$

$$a_t = \frac{\exp(d_t)}{\sum_{t=1}^{N} \exp(d_t)} \tag{4.9}$$

ここで，$\boldsymbol{v}$ は学習対象のパラメータである．

4.5.3 道徳判断ネットワーク

分散表現学習ネットワークの出力 $\boldsymbol{r}$ は道徳判断ネットワークの入力となる．それと同時に，共起情報も本ネットワークの入力として用いられる．以下で，共起情報の計算方法について述べる．

共起情報の計算

第 2 章の実験結果から，評価表現との共起情報は本タスクにおいて有用であることがわかった．そこで，本ネットワークにおいてもその共起情報を用いることを考える．共起情報の計算は以下の通りに行われる．

1. 評価表現の抽出: 評価表現辞書の中には道徳判断の際にノイズとなるような単語が多数収録されている．そこで，道徳判断の際に効果を発揮すると考えられる語彙のみを抽出し用いる．本章の研究においては第 2 章と同様のアルゴリズムが用いて評価表現の抽出が行われる．

2. 入力文からの内容語の抽出: 入力文からは内容語の原型が抽出される．第 2 章と同様に内容語として，名詞，形容詞，動詞が抽出される．

3. Web コーパスを用いた共起頻度の計算: 評価表現と内容語の共起頻度の計算が行われる．共起頻度の計算の際には，Web 日本語 N-gram が用いられる．第 2 章の時と同様に，共起頻度が取得できない場合は，単語の削除が行われる．この場合も第 2 章と同様のアルゴリズムが用いられる．

4. 共起情報の計算: 最終的に共起情報 $\boldsymbol{c}$ は以下の通りに計算される．

$$\boldsymbol{c} = \begin{cases} \boldsymbol{c'}/c'_{max} & (\boldsymbol{c'} \neq \boldsymbol{0}) \\ \boldsymbol{0} & (otherwise) \end{cases} \tag{4.10}$$

ここで，$\boldsymbol{c'}$ は各評価表現毎の共起頻度を並べたベクトルであり，c'_{max} は $\boldsymbol{c'}$ の要素の中で最大の値を表す．

道徳判断ネットワークの出力

道徳判断ネットワークの入力は，分散表現学習ネットワークで出力された $\boldsymbol{r}$ および，共起情報 $\boldsymbol{c}$ である．まず，これらの 2 つの入力が結合される．

$$\boldsymbol{e}_1 = \begin{pmatrix} \boldsymbol{r} \\ \boldsymbol{c} \end{pmatrix} \tag{4.11}$$

その後，以下のような変換を行い，出力 $\boldsymbol{y}$ が計算される．

$$\boldsymbol{e}_2 = \tanh(\boldsymbol{W}_{e_1e_2}\boldsymbol{e}_1) \tag{4.12}$$

$$\boldsymbol{y} = softmax(\boldsymbol{W}_{e_2y}\boldsymbol{e}_2) \tag{4.13}$$

ここで $softmax$ はソフトマックス関数を表す．

4.5.4 モデルの学習

上記の分散表現ネットワーク及び道徳判断ネットワークは誤差逆伝搬法を用いて学習される．誤差関数として，以下の通りに定義されるクロスエントロピー関数が用いられる．

表 4.4 道徳コーパスとして獲得されたデータの例.

Positive
親交を深める，他人と一緒に飲食する，人を信じる，人を愛する， 家で映画を楽しむ，家で映画を楽しむ，親に相談する，友達と遊ぶ， 信じて行動する，名曲を演奏する，親睦を深める，理想を求める
Negative
他人を殺す，遊びに公費を使う，嫉妬に狂う，人を憎む， 他人に呪いをかける，人を傷つける，迷惑を掛ける， 親に殺意憎悪を抱く，誇りをふみにじる，人に鞭打つ

$$loss = -\sum_{i}\sum_{j} \boldsymbol{y'}_i^j \log \boldsymbol{y}_i^j \tag{4.14}$$

この誤差関数を基に，誤差逆伝搬法により全てのパラメータの更新が行われる．

4.6 評価実験 1: 道徳コーパスの評価

4.6.1 道徳コーパスの構築

日本語 Web コーパス 2010 [131] の一部を用いて道徳コーパスの構築を行った．結果的に，278,500 文を収録したコーパスが構築された．表 4.4 に獲得されたデータの例を示す．

4.6.2 実験条件

構築された道徳コーパス中からランダムに 1,000 文 (Positive: 500 文, Negative: 500 文) を抽出した．評価の際にはまず，その文が意味を成しているか否かを判断してもらい，意味を成すと判断されたものについては以下の基準で各文を評価してもらった．

1. とても悪い

表 4.5 道徳コーパスの主観評価の結果.

	割合
ラベル付けが正しい	0.49 (487/1000)
ラベル付けが誤り	0.14 (139/1000)
文脈に依存する	0.25 (253/1000)
意味を成さない	0.12 (121/1000)

2. 悪い
3. どちらとも言えない (前後の文脈に依存する)
4. 良い
5. とても良い

4.6.3 実験結果・考察

表 4.5 に主観評価の結果を示す.

ここで, “ラベル付けが正しい” とは以下のいずれかに該当する場合である.

- 文の擬似ラベルが “Positive” であり, 主観評価の平均値が 3.30 よりも大きい場合.
- 文の擬似ラベルが “Negative” であり, 主観評価の平均値が 2.70 よりも小さい場合.

一方, “ラベル付けが誤り” とは以下のいずれかに該当する場合である.

- 文の擬似ラベルが “Negative” であり, 主観評価の平均値が 3.30 よりも大きい場合.
- 文の擬似ラベルが “Positive” であり, 主観評価の平均値が 2.70 よりも小さい場合.

また, “意味を成さない文章” かつ上記のいずれの場合にも属さない場合, “文脈に依存する” とした.

表 4.6 本実験で用いられるパラメータ.

入力層の次元数 $\boldsymbol{x'}_t$	20,642
単語の埋め込み次元数 $\boldsymbol{x}_t$	100
LSTM の次元数	100
共起情報の次元数 $\boldsymbol{c}$	40
$\boldsymbol{e}_1$ の次元数	140
$\boldsymbol{e}_2$ の次元数	50
学習回数	5,000
ミニバッチサイズ	64

4.7 評価実験 2: 擬似ラベル付きデータを用いた道徳判断精度の検証

4.7.1 データセット

データセットとして，4.6 で収集されたデータを用いた．評価用データとして 4.6.3 でラベル付けが正しいもしくはラベル付けが誤りと判断された事例の計 626 事例を用いた．但し，ラベル付けが誤りと判断された 139 事例については，付与された擬似ラベルの極性を反転させたものを正解ラベルとして用いた．

4.7.2 実験条件

表 4.6 に本実験で用いられるパラメータを示す．訓練データセットの中での出現頻度が 3 回以下の語は “未知語” として統一的に扱った．その結果，全語彙数は 20,642 となった．

共起頻度を計算する際の Positive 単語数 (n_p) および Negative 単語数 (n_n) は 20 とした．また，パラメータを更新する際の最適化手法として Adam [147] を用いた．

比較手法として以下の 5 つの手法を用いた．

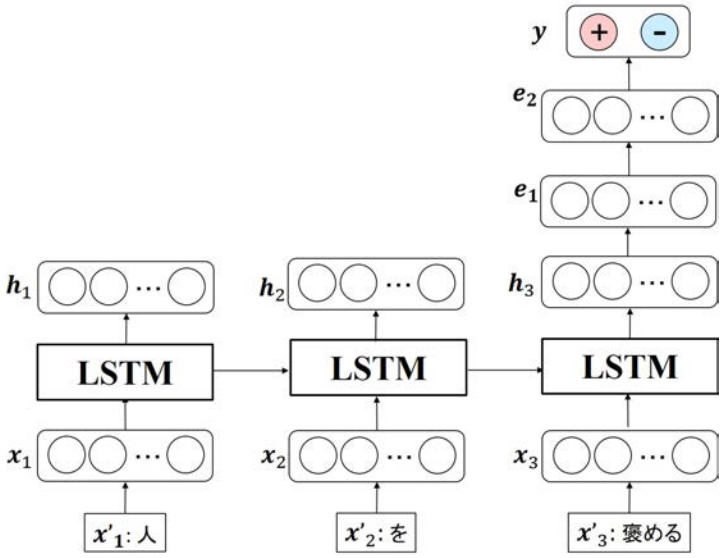

図 4.7 比較手法である Long Short-Term Memory (LSTM) のネットワーク図. 提案手法と比べ，注意機構を導入していない点及び，共起情報を使用していない点が異なる.

- 共起ベースの手法 (CO): 第 2 章で説明した，評価表現を用いた道徳判断手法である.
- ロジスティック回帰 (LR): 学習データは提案手法と同様のものを用いるが，その学習をロジスティック回帰により行う手法である，手法の詳細については，付録 C で説明している．実装は Scikit-Learn [148] を用いて行った．前処理として，ストップワードの除去や原型化などの処理は行わなかった．Positive クラスの訓練データ及び Negative クラスの訓練データそれぞれにつき，頻度上位 30,000 個の N-gram の抽出を行った．重複を削除した結果，合計で 49,547 個の N-gram 特徴量が獲得された．その内訳は unigram 特徴量が 11,282 個, bigram 特徴量が 21,256 個, trigram 特徴量が 17,009 個であった.
- Long Short-Term Memory (LSTM): 提案手法において，注意機構および共起情報を用いない場合の手法である．この手法のネットワークを図 4.7 に示す．学習の際には式 (4.14) を用いて誤差逆伝搬法によりパラメータの更新が行われる.
- 注意機構付き Long Short-Term Memory (AL): 提案手法において，共起情報を

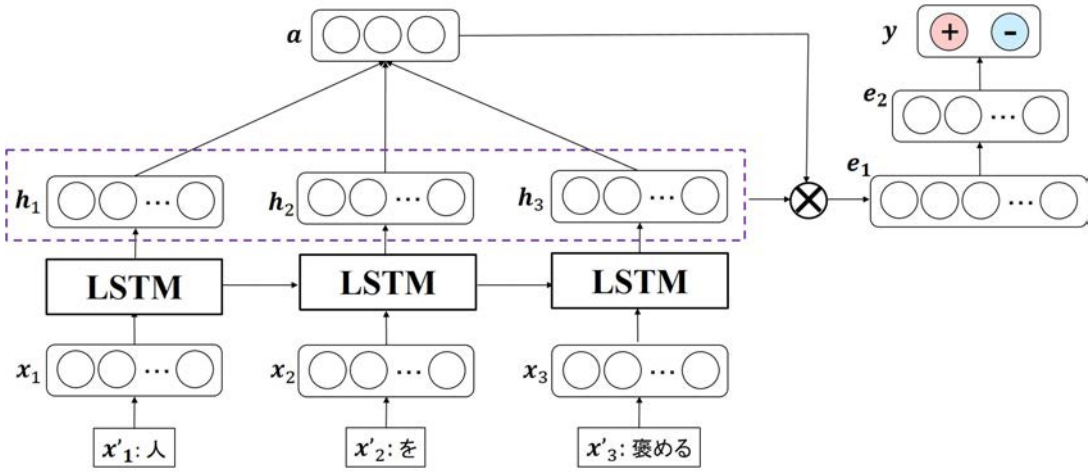

図 4.8 比較手法である注意機構付き Lpng Short-Term Memory (ALC) のネットワーク図. 提案手法と比べると，注意機構を導入している点では同様であるが，共起情報を使用していない点が異なる.

用いない場合の手法である．この手法のネットワークを図 4.8 に示す．学習の際には式 (4.14) を用いて誤差逆伝搬法によりパラメータの更新が行われる.

- 注意機構付き Long Short-Term Memory + 共起情報 (ALC): 4.5 にて説明した提案手法である.

4.7.3 実験結果・考察

道徳判断の精度

表 4.7 に各手法による道徳判断の精度を示す.
各手法間の結果を比較することにより，以下のことがわかる.

- CO vs LR: 共起ベースの手法とロジスティック回帰モデルによる結果の間には精度の向上を確認することができた．このことから，擬似ラベル付きデータを用いたことで道徳判断の精度が向上したと考えられる.

- LR vs LSTM: ロジスティック回帰モデルによる結果と LSTM による結果の間に

表 4.7 道徳判断実験の結果.

Method	Accuracy
CO	0.67 (418/626)
LR	0.80 (501/626)
LSTM	0.80 (502/626)
AL	0.82 (513/626)
ALC	0.84 (523/626)

は精度の向上を確認することはできなかった．このことから，注意機構を用いない場合の LSTM によって得られる結果は精度向上に寄与しないことがわかる．

- LSTM vs AL: LSTM により得られる結果と注意機構を用いたモデルによる結果の間には精度の向上を確認することができた．このことから，注意機構による特徴量抽出は本タスクにおいて精度向上に寄与することがわかる．

- AL vs ALC: 注意機構を用いたモデルによる結果と注意機構に加え共起情報を用いたモデルによる結果の間には精度の向上を確認することができた．このことから，共起情報を利用することにより，本タスクの精度が向上することが分かる．

文の長さの分析

図 4.9 に文の長さ毎の精度の比較を示す．ここで文の長さとは，入力文の文字数を示す．また，図中の黒い棒グラフの結果がロジスティック回帰モデルによるもの，青いものが注意機構を導入した手法によるもの，赤いものが注意機構と共起情報を導入したものを示す．

図 4.9 から，文の長さが短い場合は精度に大きな差は確認できないことが分かる．具体的には，文の長さが 7 よりも小さい場合，手法間の精度に大きな差は確認されない．一般に，文の長さが短い場合，文中に分類のために有益な情報が多く含まれることは少ないと考えられる．そのため，深層学習による手法を用いても分類のために有用な特徴量抽出を行うことができなかったと考えられる．

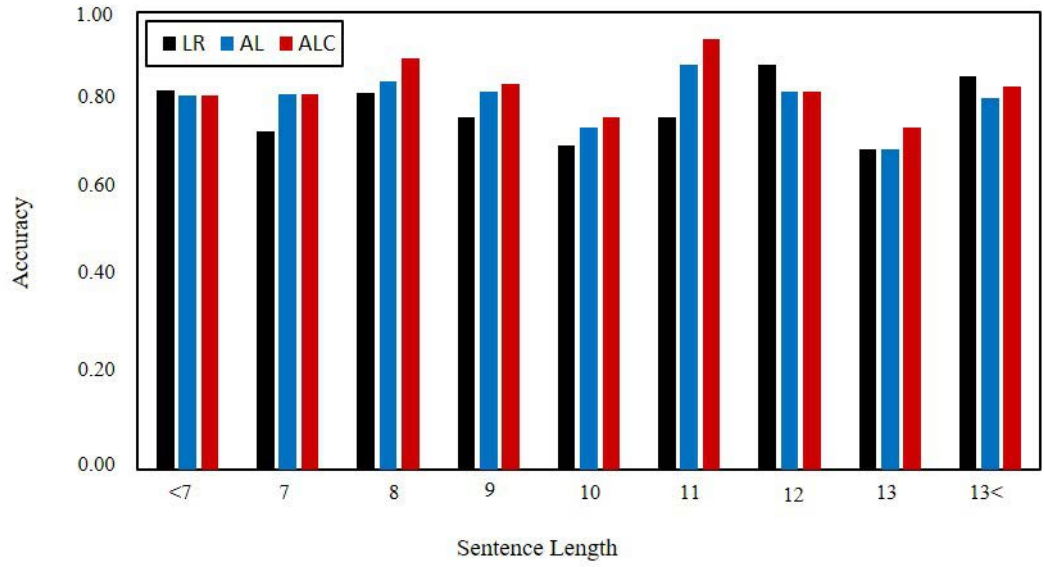

図 4.9 文の長さ毎の精度の比較.

また，文の長さが長い場合も精度に大きな差は確認できないことが分かる．一般に，文の長さが長い場合，文の長さが短い場合とは逆に分類のために有益な情報が数多く含まれることが考えられる．そのため，深層学習による手法を用いずとも，分類のために有用な特徴量が抽出できると考えられる．

最後に，文の長さが上記以外の場合 (8 以上 13 未満の場合) には精度に差が確認できる．これは，文の長さが短文，長文ではない場合，特徴量をどのように抽出するかによって大きな差が生じるためであると考えられる．注意機構の導入及び，共起情報の考慮によって，分類のために有益な情報を抽出することが可能になり，結果的に精度が向上したと考えられる．

定性的分析

定量的な分析に続いて，定性的な分析を行った．分析の際には式 (4.9) により得られる $\boldsymbol{a}$ を用いた．図 4.10 に注意機構の重みパラメータ $\boldsymbol{a}$ の可視化を示す．

図 4.10 には 4 つの例が示されており，色が濃いほどパラメータの値が大きいことを意味する．例えば,「(a) 人を愛する」という文では,「愛する」という単語の情報を基

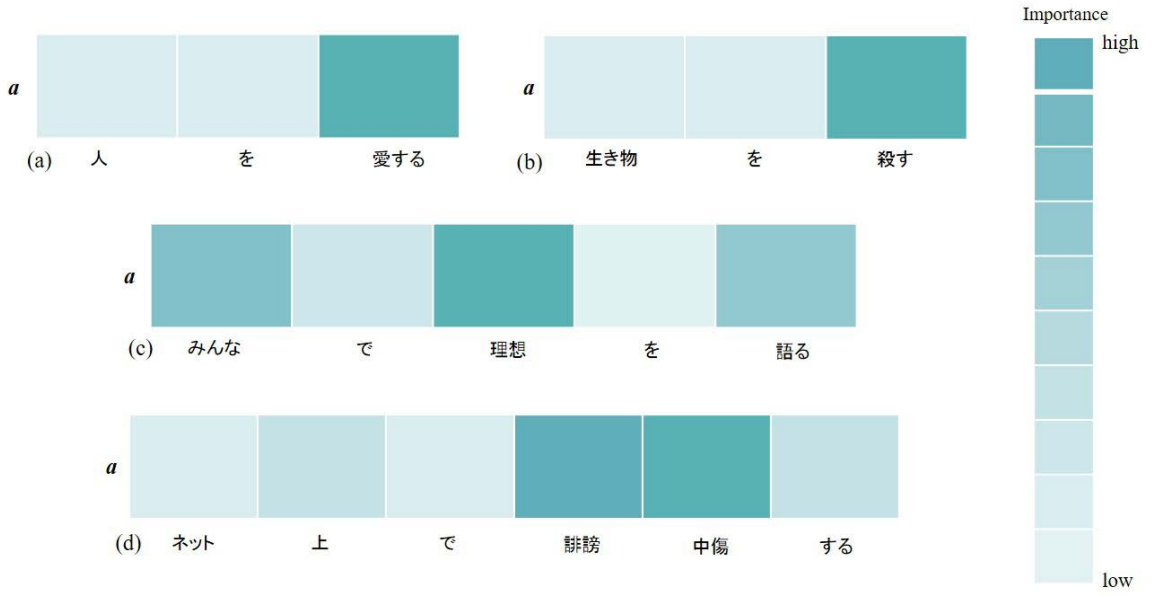

図 4.10 注意機構の可視化. 色が濃いほど注意機構の値が大きいことを表し，モデルの出力に大きな影響を与えていると考えられる．例えば，“(a) 人を愛する” という例では “愛する” という表現に大きな影響が与えられていることが分かる．

に道徳判断が行なわれていることが分かる．同様にして,「(b) 生き物を殺す」という文では,「殺す」という単語の情報により大きな値が割り当てられていることが分かる．

一方，判断の際に重要となる単語が複数ある場合，それらをまとめて検知できることが (c), (d) の例から分かる．例えば,「(c) みんなで理想を語る」という文では,「みんな」,「理想」,「語る」という単語の情報に対応する注意機構のパラメータの値が大きくなっている．また,「(d) ネット上で誹謗中傷する」という文では,「誹謗」,「中傷」という単語に大きな注意が張られている．

このようにして，注意機構のパラメータを可視化することにより，どの部分を基にその判断が行なわれているかの分析が可能となる．一般的に深層学習では内部の分析が難しいため，モデルが何故その出力をしたのかを理解することは困難であると言われている．しかし，本研究のような注意機構を用いることで，ただ精度を向上するだけではなく，出力の根拠を可視化することが可能となることが分かる．

4.8 本章で提案した手法のまとめ

本章では，自動獲得されたラベル付きデータを用いた道徳判断手法について述べた．提案手法ではまず，擬似ラベル付きデータの自動獲得が行われる．次に獲得されたデータを用いて，機械学習モデルが構築される．

機械学習モデルには，深層学習によるモデルを提案した．具体的には，注意機構を有する LSTM による特徴量の抽出と，共起情報の統合が本研究の大きな特長である．

評価実験の結果，以下のことが確認された．

- 擬似ラベル付きデータを用いることによる精度の向上: 共起ベースの手法に比べ擬似ラベル付きデータを用いた手法による精度向上が確認された．このことから，擬似ラベル付きデータを獲得する手法の有効性が示唆された．

- 深層学習導入による精度の向上：注意機構付き LSTM の導入によって，N-gram 特徴量よりも良い特徴量が獲得されたことが示唆された．また，注意機構のパラメータを可視化することで，道徳判断の際にどこに着目して出力が行なわれているかの分析が可能となった．

- 共起情報と言語情報を統合することによる精度の向上：共起情報を考慮することで更に精度が向上することが確認された．

第 5 章

結論

5.1 本研究のまとめ

本論文では，Web 上のテキストデータを用いた道徳的な知識の自動獲得について述べた．Web 上には多種多様なデータが大量に存在するため，自然言語処理技術を適用することで様々な事例に対して道徳判断が可能になった．

第 2 章では，評価表現を用いた道徳判断手法を提案した．具体的には，入力文と評価表現の共起情報を基にした手法の提案を行った．このとき，あらかじめ構築されている評価表現辞書の中には道徳判断の際にノイズとなるような語彙も含まれている．そこで，評価表現辞書の中から，道徳判断の際に効果を発揮する語彙を抽出する手法を提案した．また，共起情報が取れない場合が存在する問題に対し，文中の非重要単語を削除する手法を提案した．

第 3 章では，対象言語を英語に変更した手法の提案を行った．第 2 章の手法をそのまま適応することは困難であると考えられたため，分散表現を用いた道徳判断手法を提案した．一般的な手法で学習された分散表現は様々なタスクで精度が高いことが報告されている一方で，学習の性質上，Positive な単語と Negative な単語の類似度が高くなってしまう．そのため，道徳判断の際にそのままの分散表現を使うことは精度の面から望ましくないと考えられる．この問題に対し，分散表現を道徳判断タスク用に再学習する手法を提案した．

第 4 章では，擬似ラベル付きデータを用いた道徳判断手法を提案した．まず，構文情報，接続表現，評価表現を用いて，擬似ラベル付きデータを自動で獲得する手法について述べた．次に，自動獲得されたデータを訓練データとして，教師あり学習を行う手法について説明した．具体的には，注意機構付き Long Short-Term Memory で教師

あり学習を行うことでより高い精度で道徳判断が可能であることを示した．それと同時に，言語の表層情報だけではなく，共起情報を追加特徴量とすることで更に高い精度で道徳判断が可能であることを示した．また，注意機構パラメータの可視化によって，判断の際に重視された箇所の分析が可能となった．このように事例ベースで道徳判断を行うことで，共起ベースの手法では困難であった事例に対しても適切な道徳判断が可能となった．

上記で提案した手法はいずれも Web 上のテキストデータを用いており，その量が多くなればなるほど，精度は上がっていくと考えられる．Web 上のテキストデータは毎年増加しており，今後より多くのデータが集まることが予想されるため，その意義は大きくなっていくものと考えられる．

5.2 今後の課題

本章では，今後の課題について述べる．大きく分けて，道徳判断手法自体の課題と将来の展望について述べる．

5.2.1 道徳判断手法の高度化

ノイズの存在するデータの取り扱い

第 4 章で自動獲得されたデータの中にはノイズとなるような事例も存在する．具体的には，第 4 章での評価実験の結果，正しくラベル付けされたデータは 5 割弱という結果になった．一般的に教師あり学習のモデル構築では，人手により正しくラベル付けされているデータを取り扱うため，本研究のようなノイズデータが存在することは前提とはされていない．

このようなノイズデータが存在する状況下での機械学習については，近年様々な分野で盛んに研究が行われている．例えば，関係抽出の分野では本研究と同じようにノイズ付きのデータを訓練データとして用いるため，それを適切にモデル化する研究が行われている．自然言語処理分野に限らず，画像認識の分野でもこのようなノイズ付きデータを前提とした研究が存在する．

上記で提案されているモデルは，本研究においても適用可能であり，適切にモデルを構築することで精度が向上すると考えられる．このノイズが存在するデータの取り扱いというトピックは道徳判断分野のみならず，機械学習の理論的側面からも興味深い研究内容であると考えられる．

5.2.2 将来の展望

実世界とのグラウンディング

本研究では主にテキストを入力とし，その文が示す行動に対する善悪性を出力としている．そのため，テキストデータで書かれた行動が実世界でどういった行動を示すのかについて紐づけられているわけではない．将来的にロボットの行動指針として活用することを考えると，本研究のような入力文が実世界のどういった動作と結びつくのかについての知識が必要であると考えられる．

こういった，言語情報と実世界のグラウンディングは近年研究が盛んになっている分野でもある．今後どのような方向で技術が進歩していくかは自明ではないが，このようなセンサーや画像，音声といったようなマルチモーダルな情報を結びつける研究が進展することはほぼ間違いないと考えられる．技術の進歩の中で，テキストに書かれた行動と実世界での動作の対応付けを行うことはロボット倫理学としても重要なことであると考えられる．

期待される応用可能性

本研究では，人工知能が言語理解や状況判断を行うために必要となる道徳的な常識の獲得を行った．今後，人間とインタラクションを行い，自ら判断を行う人工知能は増えていくと考えられる．そのとき，言語理解，適切な応答，価値判断の根拠として人間が保持している知識を考慮することは非常に重要である．特に道徳性については，今までの研究であまり考慮されておらず，本研究で獲得された知識は大変重要なものになると考えられる．具体的には大きく分けて 2 つの応用が考えられる．

1 つ目は, 自然言語処理システムの知識源として利用することである．既に, WordNet などの単語単位の知識を用いて自然言語処理システムの性能を向上させる研究が存在

する．例えばYangらは, 知識を使用した言語処理のための汎用的な手法を提案している [149]．本研究で獲得された道徳的な常識も, 同じような方向性で自然言語処理システムの性能を向上させるために用いることが可能である．例えば, 人間とインタラクションを行う対話システムなどへの応用が考えられる．

応用先の 2 つ目は, ロボットなどの自律型人工知能の知識として利用することである．実際に，Web から獲得された知識が, ロボットの行動の価値判断の際に役立つことを示した研究が存在する．Takagi らは，Web 上で獲得された知識が掃除ロボットの価値判断において有用であることを示した [150]．彼らの手法では,「赤ちゃんが寝ているような状況では音を立てることが善くない」ことであるという知識を獲得することができる．そのため，部屋が汚れているような状態でも，赤ちゃんが寝ているのであれば，掃除を行わないというような価値判断を行うことが可能となる．

このようなシステムの社会実装のためには, 様々な分野の知見が必要になると考えられる．上述した実世界とのグラウンディングのためには, 言語情報だけではなく画像, 音声, センサーなどの様々な情報の処理が必要となってくる．また, 本研究では実世界の情報を一度言語情報に変換し, その情報を基に善悪判断を行うことを想定しているが, 言語情報を使用せずに善悪判断を行うことも考えられる．そのような手法とは相補的な関係にあり, 両者の手法を結合することでより精度の高い判断が可能になることが予想される．

工学分野だけではなく, 倫理学, 社会科学, 心理学などの, 道徳に関する研究が盛んな分野の知見も重要である．例えば,Voiklis らは, 人間に期待される倫理感とロボットに期待される倫理感は異なることを心理学的な実験により示している [151]．このような研究成果は, 社会実装を行う上で大変貴重なものであると考えられる．

参考文献

[1] “人工知能学会 倫理指針,” http://ai-elsi.org/wp-content/uploads/2017/02/人工知能学会倫理指針.pdf.

[2] “ASILOMAR AI PRINCIPLES,” https://futureoflife.org/ai-principles/.

[3] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, Vol. 9, No. 8, pp. 1735–1780, 1997.

[4] “道徳,” https://ja.wikipedia.org/wiki/道徳.

[5] J. Weizenbaum, “ELIZA—a computer program for the study of natural language communication between man and machine,” *Communications of the ACM*, Vol. 9, No. 1, pp. 36–45, 1966.

[6] J. McCarthy and P. J. Hayes, “Some Philosophical Problems from the Standpoint of Artificial Intelligence,” In *Machine Intelligence*, pp. 463–502. Edinburgh University Press, 1969.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” In *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[8] 松原 仁, “あから2010勝利への道特集,” 情報処理学会論文誌, Vol. 52, No. 2, pp. 152–190, 2011.

[9] “コンピュータ将棋プロジェクトの終了宣言,” http://www.ipsj.or.jp/50anv/shogi/20151011.html.

[10] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al., "Mastering the game of go with deep neural networks and tree search," *Nature*, Vol. 529, No. 7587, pp. 484–489, 2016.

[11] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. Lillicrap, F. Hui, L. Sifre, G. van den Driessche, T. Graepel, and D. Hassabis, "Mastering the game of go without human knowledge," *Nature*, Vol. 550, pp. 354–359, 2017.

[12] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, et al., "Mastering chess and shogi by self-play with a general reinforcement learning algorithm," *arXiv preprint arXiv:1712.01815*, 2017.

[13] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[14] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," *arXiv preprint arXiv:1508.04025*, 2015.

[15] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, et al., "Google's neural machine translation system: Bridging the gap between human and machine translation," *arXiv preprint arXiv:1609.08144*, 2016.

[16] "人工知能の13歳の少年、チューリングテストに"合格"," `http://www.itmedia.co.jp/news/articles/1406/09/news049.html`.

[17] "テスラのオートパイロット（自動運転支援機能）の発表内容詳細," `http://blog.evsmart.net/ev-news/tesla-autopilot-2-0/`.

[18] "トロッコ問題," `https://ja.wikipedia.org/wiki/トロッコ問題`.

[19] “爆破ロボットで容疑者殺害,” https://mainichi.jp/articles/20160709/k00/00e/030/306000c.

[20] “『Google フォト』が黒人２人の写真を“ゴリラ”と自動認識して物議,” https://rocketnews24.com/2015/07/06/604510/.

[21] “Tay (人工知能),” https://ja.wikipedia.org/wiki/Tay_(人工知能).

[22] “ロボット倫理学,” https://ja.wikipedia.org/wiki/ロボット倫理学.

[23] 松尾 豊, 西田 豊明, 堀 浩一, 武田 英明, 長谷 敏司, 塩野誠, 服部 宏充, “人工知能学会倫理委員会の取組み,” 人工知能, Vol. 30, No. 3, pp. 358–364, 2015.

[24] “人工知能と人間社会に関する懇談会,” http://www8.cao.go.jp/cstp/tyousakai/ai/.

[25] “ロボット倫理憲章,” https://ja.wikipedia.org/wiki/ロボット倫理憲章.

[26] S. Bringsjord, K. Arkoudas, and P. Bello, “Toward a general logicist methodology for engineering ethically correct robots,”, Vol. 21, No. 4, pp. 38–44, 2006.

[27] T. M. Powers, “Prospects for a kantian machine,” *IEEE Intelligent Systems*, Vol. 21, No. 4, pp. 46–51, 2006.

[28] L. M. Pereira and A. Saptawijaya, “Modelling morality with prospective logic,” *International Journal of Reasoning-based Intelligent Systems*, Vol. 1, No. 3-4, pp. 209–221, 2009.

[29] J. Fleetwood, W. Vaught, D. Feldman, E. Gracely, Z. Kassutto, and D. Novack, “Medethex online: a computer-based learning program in medical ethics and communication skills,” *Teaching and Learning in Medicine*, Vol. 12, No. 2, pp. 96–104, 2000.

[30] B. M. McLaren and K. D. Ashley, “Case-based comparative evaluation in truth-teller,” In *Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society*, 1995.

[31] B. McLaren, "Extensionally defining principles and cases in ethics: an ai model," *Artificial Intelligence Journal*, Vol. 150, pp. 145–181, 2003.

[32] 笠原 要, 稲子 希望, 加藤 恒明, "テキストデータを用いた類義語の自動作成," 人工知能学会論文誌, Vol. 4, No. 18, pp. 221–232, 2003.

[33] 渡部 啓吾, D. Bollegala, 松尾 豊, 石塚 満, "検索エンジンを用いた関連語の自動抽出," 知能と情報 : 日本知能情報ファジィ学会誌 : journal of Japan Society for Fuzzy Theory and Intelligent Informatics, Vol. 23, No. 5, pp. 739–748, 2011.

[34] R. Schwartz, R. Reichart, and A. Rappoport, "Symmetric pattern based word embeddings for improved word similarity prediction," In *Proceedings of CoNLL*, 2015.

[35] 城光 英彰, 松田 源立, 山口 和紀, "文脈限定 Skip-gram による同義語獲得," 自然言語処理学会論文誌, Vol. 24, No. 2, pp. 187–204, 2017.

[36] 内海 慶, 小町 守, "ウェブ検索クエリログとクリックスルーログを用いた同義語獲得," 情報処理学会論文誌, Vol. 6, No. 1, pp. 16–28, 2013.

[37] C. N. dos Santos, B. Xiang, and B. Zhou, "Classifying relations by ranking with convolutional neural networks," In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pp. 626–634, 2015.

[38] M. Miwa and M. Bansal, "End-to-end relation extraction using lstms on sequences and tree structures," In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 1105–1116, 2016.

[39] L. Wang, Z. Cao, G. de Melo, and Z. Liu, "Relation classification via multi-level attention cnns," In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 1298–1307, 2016.

[40] M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni, "Open information extraction from the web," In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, pp. 2670–2676, 2007.

[41] K. Eichler and H. Hemsen, "Unsupervised relation extraction from web documents," In *Proceedings of the 6th International Language Resources and Evaluation (LREC)*, 2008.

[42] M. Mintz, S. Bills, R. Snow, and D. Jurafsky, "Distant supervision for relation extraction without labeled data," In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pp. 1003–1011, 2009.

[43] S. Riedel, L. Yao, and A. McCallum, "Modeling relations and their mentions without labeled text," In *Proceedings of the 2010 European Conference on Machine Learning and Knowledge Discovery in Databases: Part III*, pp. 148–163, 2010.

[44] R. Hoffmann, C. Zhang, X. Ling, L. Zettlemoyer, and D. S. Weld, "Knowledge-based weak supervision for information extraction of overlapping relations," In *ACL-HLT 2011 - Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 541–550, 2011.

[45] D. Zeng, K. Liu, Y. Chen, and J. Zhao, "Distant supervision for relation extraction via piecewise convolutional neural networks," In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1753–1762, 2015.

[46] G. Ji, K. Liu, S. He, and J. Zhao, "Distant supervision for relation extraction with sentence-level attention and entity descriptions," In *Proceedings of The Thirty-First AAAI Conference on Artificial Intelligence*, pp. 3060–3066, 2017.

[47] 進 義治, 黒橋 禎夫, "名詞関連語知識に基づく文章のグラフ表現とその応用," 言語処理学会 第 20 回年次大会 発表論文集, pp. 1007–1010, 2014.

[48] 町田 雄一郎, 河原 大輔, 黒橋 禎夫, 颯々野 学, “関連語知識獲得のための対話システム上の連想ゲームのデザイン,” 情報処理学会論文誌, Vol. 57, No. 3, pp. 1058–1068, 2016.

[49] 大谷 直樹, 河原 大輔, 黒橋 禎夫, 鍜治 伸裕, 颯々野 学, “連想ゲームによるコモンセンス知識の獲得,” 言語処理学会 第 22 回年次大会 発表論文集, pp. 897–900, 2016.

[50] G. A. Miller, “Wordnet: a lexical database for english,” *Communications of the ACM*, Vol. 38, No. 11, pp. 39–41, 1995.

[51] D. B. Lenat, G. A. Miller, and T. Yokoi, “Cyc, WordNet, and EDR: critiques and responses,” *Communications of the ACM*, Vol. 38, No. 11, pp. 45–48, 1995.

[52] D. Lenat, “Cyc: A large-scale investment in knowledge infrastructure,” *Communications of the ACM*, Vol. 38, No. 11, pp. 33–38, 1995.

[53] P. Singh, T. Lin, E. T. Mueller, G. Lim, T. Perkins, and W. L. Zhu, “Open Mind Common Sense: Knowledge Acquisition from the General Public,” In *Proceeding On the Move to Meaningful Internet Systems*, pp. 1223–1237, 2002.

[54] H. Liu and P. Singh, “ConceptNet &Mdash; A Practical Commonsense Reasoning Tool-Kit,” *BT Technology Journal*, Vol. 22, No. 4, pp. 211–226, 2004.

[55] F. Bond, H. Isahara, S. Fujita, K. Uchimoto, T. Kuribayashi, and K. Kanzaki, “Enhancing the japanese wordnet,” In *Proceedings of the 7th Workshop on Asian Language Resources*, pp. 1–8, 2009.

[56] 荻野 孝野, “EDR 電子化辞書について,” 情報処理学会情報メディア研究会, Vol. 34, No. 7, pp. 31–38, 1998.

[57] “日本語語彙大系 CD-ROM 版,” `http://www.kecl.ntt.co.jp/mtg/resources/GoiTaikei/`.

[58] 白井 諭, 大山 芳史, 池原 悟, 宮崎 正弘, 横尾 昭男, “日本語語彙大系について,” 情報処理学会研究報告, Vol. 98, pp. 47–52, 1998.

[59] 岡本 潤, 石崎 俊, “概念間距離の定式化と既存電子化辞書との比較,” 自然言語処理, Vol. 8, No. 4, pp. 37–54, 2001.

[60] 玉川 奨, “日本語 Wikipedia オントロジーの自動構築と評価,” PhD thesis, 慶應義塾大学, 2013.

[61] D. Lin and P. Pantel, “Discovery of inference rules for question-answering,” *Natural Language Engineering*, Vol. 7, No. 4, p. 343?360, 2001.

[62] 乾 孝司, 乾 健太郎, 松本 裕治, “接続標識「ため」に基づく文書集合からの因果関係知識の自動獲得,” 情報処理学会論文誌, Vol. 45, No. 3, pp. 919–932, 2004.

[63] 阿部 修也, 乾 健太郎, 松本 裕治, “項の共有関係と統語パターンを用いた事態間関係獲得,” 自然言語処理学会論文誌, Vol. 45, No. 3, pp. 121–139, 2010.

[64] H. J. Levesque, “The winograd schema challenge,” In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*, pp. 552–561, 2011.

[65] N. Inoue and K. Inui, “ILP-Based Reasoning for Weighted Abduction,” In *Plan, Activity, and Intent Recognition*, pp. 25–32, 2011.

[66] A. Rahman and V. Ng, “Resolving Complex Cases of Definite Pronouns: The Winograd Schema Challenge,” In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 777–789, 2012.

[67] A. Sharma, N. H. Vo, S. Aditya, and C. Baral, “Towards Addressing the Winograd Schema Challenge - Building and Using a Semantic Parser and a Knowledge Hunting Module,” In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1319–1325, 2015.

[68] R. C. Schank and R. P. Abelson, "Scripts, Plans, Goals and Understanding: an Inquiry into Human Knowledge Structures," L. Erlbaum, 1977.

[69] M. Regneri, A. Koller, and M. Pinkal, "Learning script knowledge with web experiments," *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 979–988, 2010.

[70] B. Jans, S. Bethard, I. Vulić, and M. F. Moens, "Skip n-grams and ranking functions for predicting script events," In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 336–344, 2012.

[71] 乾 孝司, 奥村 学, "テキストを対象とした評価情報の分析に関する研究動向," 自然言語処理, Vol. 13, No. 3, pp. 201–241, 2006.

[72] 小林 のぞみ, 乾 健太郎, 松本 裕治, 立石 健二, 福島 俊一, "意見抽出のための評価表現の収集," 自然言語処理, Vol. 12, No. 2, pp. 203–222, 2005.

[73] 小林 のぞみ, 乾 健太郎, 松本 裕治, "意見情報の抽出/構造化のタスク仕様に関する考察," 情報処理学会研究報告 NL171-18, pp. 111–118, 2006.

[74] H. Takamura, T. Inui, and M. Okumura, "Extracting Semantic Orientations of Words using Spin Model," In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 133–140, 2005.

[75] 高村 大也, 乾 孝司, 奥村 学, "スピンモデルによる単語の感情極性抽出," 情報処理学会論文誌, Vol. 47, No. 2, pp. 627–637, 2006.

[76] N. Kaji and M. Kitsuregawa, "Building Lexicon for Sentiment Analysis from Massive HTML Documents," In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-CoNLL2007)*, pp. 1075–1083, 2007.

[77] 佐野 大樹, “日本語における評価表現の分類体系 : アプレイザル理論をベースに,” 電子情報通信学会技術研究報告. NLC, 言語理解とコミュニケーション, Vol. 110, No. 400, pp. 19–24.

[78] “筑波大学文単位評価極性タグ付きコーパス（TSUKUBA コーパス）,” https://rit.rakuten.co.jp/data_release_ja/.

[79] N. Kaji and M. Kitsuregawa, “Automatic construction of polarity-tagged corpus from html documents,” In *Proceedings of ACL/COLING*, pp. 452–459, 2006.

[80] 鍛冶 伸裕, 喜連川 優, “HTML 文書集合からの評価文の自動収集,” 自然言語処理, Vol. 15, No. 3, pp. 77–90, 2008.

[81] “京都観光ブログの評価情報付与データ,” https://alaginrc.nict.go.jp/resources/nict-resource/li-info/li-outline.html#A-10.

[82] P. Turney, “Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews,” In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pp. 417–424, 2002.

[83] X. Wan, “Co-training for cross-lingual sentiment classification,” In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pp. 235–243, 2009.

[84] S. Dasgupta and V. Ng, “Mine the easy, classify the hard: A semi-supervised approach to automatic sentiment classification,” In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pp. 701–709, 2009.

[85] S. Li, C.-R. Huang, G. Zhou, and S. Y. M. Lee, “Employing Personal/Impersonal Views in Supervised and Semi-supervised Sentiment Classification,” In *Proceed-*

ings of the 48th Annual Meeting of the Association for Computational Linguistics, pp. 414–423, 2010.

[86] P. Biyani, C. Caragea, P. Mitra, C. Zhou, J. Yen, G. E. Greer, and K. Portier, "Co-training over Domain-independent and Domain-dependent Features for Sentiment Analysis of an Online Cancer Support Community," In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 413–417, 2013.

[87] M. Yang, W. Tu, Z. Lu, W. Yin, and K. Chow, "LCCT: A Semi-supervised Model for Sentiment Classification," In *Proceedings of The 2015 Annual Conference of the North American Chapter of the ACL (NAACL)*, pp. 546–555, 2015.

[88] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment Classification using Machine Learning Techniques," In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 79–86, 2002.

[89] B. Pang and L. Lee, "Seeing Stars: Exploiting Class Relationships For Sentiment Categorization With Respect To Rating Scales," In *Proceedings of the 43rd Annual Meeting of the ACL*, pp. 115–124, 2005.

[90] S. Matsumoto, H. Takamura, and M. Okumura, "Sentiment classification using word sub-sequences and dependency sub-trees," In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 301–311. Springer, 2005.

[91] R. Socher, J. Pennington, E. H. Huang, A. Y. Ng, and C. D. Manning, "Semi-supervised Recursive Autoencoders for Predicting Sentiment Distributions," In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 151–161, 2011.

[92] Q. Qian, B. Tian, M. Huang, Y. Liu, X. Zhu, and X. Zhu, "Learning Tag Embeddings and Tag-specific Composition Functions in Recursive Neural Network," In *Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1365–1374, 2015.

[93] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts, "Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank," In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1631–1642, 2013.

[94] D. Tang, B. Qin, and T. Liu, "Document Modeling with Gated Recurrent Neural Network for Sentiment Classification," In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1422–1432, 2015.

[95] Y. Wang, M. Huang, L. Zhao, and X. Zhu, "Attention-based lstm for aspect-level sentiment classification," In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 606–615, 2016.

[96] K. S. Tai, R. Socher, and C. D. Manning, "Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks," In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pp. 1556–1566, 2015.

[97] C. dos Santos and M. Gatti, "Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts," In *Proceedings of the 25th International Conference on Computational Linguistics (COLING)*, 2014.

[98] Y. Kim, "Convolutional Neural Networks for Sentence Classification," In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1746–1751, 2014.

[99] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision,", Vol. 1, No. 12, pp. 1–6, 2009.

[100] J. Deriu, M. Gonzenbach, F. Uzdilli, Lucchi Aur?lien, V. D. Luca, M. Jaggi, "swisscheese at semeval-2016 task 4: Sentiment classification using an ensemble

of convolutional neural networks with distant supervision," In *Proceedings of the 10th International Workshop on Semantic Evaluation*, pp. 1124–1128, 2016.

[101] S. Müller, T. Huonder, J. Deriu, and M. Cieliebak, "TopicThunder at SemEval-2017 Task 4: Sentiment Classification Using a Convolutional Neural Network with Distant Supervision," In *Proceedings of the 11th International Workshop on Semantic Evaluation*, pp. 766–770, 2017.

[102] M. Minsky, "ミンスキー博士の脳の探検ー常識・感情・自己とはー," 共立出版, 2009.

[103] D. B. Lenat, "Cyc: A large-scale investment in knowledge infrastructure," *Commun.ACM*, Vol. 38, No. 11, pp. 33–38, 1995.

[104] P. Singh, "The open mind common sense project.," *KurzweilAI.net*, 2002.

[105] C. Havasi, J. Pustejovsky, R. Speer, and H. Lieberman, "Applying common sense using dimensionality reduction.," *Inteligent Systems*, Vol. 24, No. 4, pp. 24–35, 2009.

[106] R. Speer and H. Liberman, "Analogyspace : Reducing the dimensionality of common sense knowledge.," *Artificial Intelligence*, Vol. 22, No. 4, pp. 548–553, 2008.

[107] D. B. Lenat, G. A. Miller, and T. Yokoi, "Cyc, wordnet, and edr: critiques and responses," *Communications of the ACM*, Vol. 38, No. 11, pp. 45–48, 1995.

[108] C. Elkan and R. Greiner, "Building large knowledge-based systems: representation and inference in the cyc project.," *Artificial Intelligence*, Vol. 61, No. 1, pp. 41–52, 2006.

[109] P. Danielson, "Artificial Morality: Virtuous Robots for Virtual Games," Routledge, 1992.

[110] C. Allen, G. Varner, and J. Zinser, "Prolegomena to any future artificial moral agent," *Journal of Experimental and Teoretical Artificial Intelligence*, Vol. 12, No. 3, pp. 251–261, 2000.

[111] L. Floridi and J. W. Sanders, "On the morality of artificial agents," *Mind and Machine*, Vol. 14, No. 3, pp. 349–379, 2004.

[112] P. Lichocki, A. Billard, and P. Kahn, "The ethical landscape of robotics," *Robotics , Automation Magazine, IEEE*, Vol. 18, No. 1, pp. 39–50, 2011.

[113] W. Wallach and C. Allen, "Moral Machines: Teaching Robots Right from Wrong," Oxford University Press, 2009.

[114] M. Anderson, S. Anderson, and C. Armen, "An approach to computing ethics," *Intelligent Systems, IEEE*, Vol. 21, No. 4, pp. 56–63, 2006.

[115] D. Gllies, "Artificial Intelligence and Sientific Method," Oxford University Press, 1999.

[116] C. Allen, G. Varner, and J. Zinser, "Prolegomena to any future artificial moral agent," *Journal of Experimental and Theoretical Artificial Intelligence*, Vol. 12, No. 3, pp. 251–261, 2000.

[117] T. Kudo, K. Yamamoto, and Y. Matsumoto, "Applying conditional random fields to Japanese morphological analysis," In *Proceedings of the 2004 conference on empirical methods in natural language processing*, 2004.

[118] Y. Sawai and K. Yamamoto, "Estimating level of public interest for documents," *Journal of natural language processing*, Vol. 15, No. 2, pp. 101–136, 2008.

[119] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," In *Advances in neural information processing systems*, pp. 3111–3119, 2013.

[120] Y. Miyao and J. Tsujii, "Feature forest models for probabilistic HPSG parsing," *Computational linguistics*, Vol. 34, No. 1, pp. 35–80, 2008.

[121] "The british national corpus, version 3 (bnc xml edition) 2007," `http://www.natcorp.ox.ac.uk/`.

[122] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

[123] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *arXiv preprint arXiv:1607.04606*, 2016.

[124] O. Levy and Y. Goldberg, "Neural word embedding as implicit matrix factorization," In *Advances in neural information processing systems*, pp. 2177–2185, 2014.

[125] A. Neelakantan, J. Shankar, A. Passos, and A. McCallum, "Efficient non-parametric estimation of multiple embeddings per word in vector space," *arXiv preprint arXiv:1504.06654*, 2015.

[126] D. Tang, F. Wei, N. Yang, M. Zhou, T. Liu, and B. Qin, "Learning sentiment-specific word embedding for twitter sentiment classification," In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1, pp. 1555–1565, 2014.

[127] R. Das, M. Zaheer, and C. Dyer, "Gaussian LDA for topic models with word embeddings," In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Vol. 1, pp. 795–804, 2015.

[128] C.-C. Lin, W. Ammar, C. Dyer, and L. Levin, "Unsupervised pos induction with word embeddings," *arXiv preprint arXiv:1503.06760*, 2015.

[129] M. Bansal, K. Gimpel, and K. Livescu, "Tailoring continuous word representations for dependency parsing," In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 809–815, 2014.

[130] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky, "The stanford corenlp natural language processing toolkit," In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pp. 55–60, 2014.

[131] "日本語 Web コーパス 2010," `http://s-yata.jp/corpus/nwc2010/`.

[132] "Word2vec," `https://code.google.com/p/word2vec/`.

[133] 徳久良子 , 乾健太郎 , 松本裕治 , "Web から獲得した感情生起要因コーパスに基づく感情推定," 情報処理学会論文誌, Vol. 50, No. 4, pp. 1365–1374, 2009.

[134] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," In *Advances in neural information processing systems*, pp. 3104–3112, 2014.

[135] P. Li, W. Lam, L. Bing, and Z. Wang, "Deep recurrent generative decoder for abstractive text summarization," *arXiv preprint arXiv:1708.00625*, 2017.

[136] W. Wang, N. Yang, F. Wei, B. Chang, and M. Zhou, "Gated self-matching networks for reading comprehension and question answering," In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1, pp. 189–198, 2017.

[137] F. Seide, G. Li, and D. Yu, "Conversational speech transcription using context-dependent deep neural networks," In *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[138] P. Wang, Y. Qian, F. K. Soong, L. He, and H. Zhao, "Part-of-speech tagging with bidirectional long short-term memory recurrent neural network," *arXiv preprint arXiv:1510.06168*, 2015.

[139] D. Chen and C. Manning, "A fast and accurate dependency parser using neural networks," In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 740–750, 2014.

[140] "りんな," https://ja.wikipedia.org/wiki/りんな_(人工知能).

[141] "IBM watson visual recognition," https://www.ibm.com/watson/jp-ja/developercloud/visual-recognition.html#try-it-out.

[142] "IBM watson speech to text," https://www.ibm.com/watson/jp-ja/developercloud/speech-to-text.html#try-it-out.

[143] "IBM watson personaliti insights," https://www.ibm.com/watson/jp-ja/developercloud/personality-insights.html.

[144] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," In *Computer Vision and Pattern Recognition (CVPR)*, pp. 3156–3164. IEEE, 2015.

[145] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," In *International Conference on Machine Learning*, pp. 2048–2057, 2015.

[146] T. Kudo and Y. Matsumoto, "Japanese dependency analysis using cascaded chunking," In *CoNLL 2002: Proceedings of the 6th Conference on Natural Language Learning 2002 (COLING 2002 Post-Conference Workshops)*, pp. 63–69, 2002.

[147] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[148] "Scikit-learn: Machine learning in python," `http://scikit-learn.org/stable/index.html`.

[149] B. Yang and T. Mitchell, "Leveraging knowledge bases in lstms for improving machine reading," In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1436–1446, 2017.

[150] K. Takagi, R. Rzepka, and K. Araki, "Just keep tweeting, dear: Web-mining methods for helping a social robot understand user needs," In *Symposium of Association for the Advancement of Artificial Intelligence (AAAI) Spring Symposia*, pp. 60–65, 2011.

[151] J. Voiklis, B. Kim, C. Cusimano, and B. F. Malle, "Moral judgments of human vs. robot agents," In *25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pp. 775–780, 2016.

謝辞

本研究は，著者が慶應義塾大学大学院理工学研究科後期博士課程在学中に，同大学理工学部萩原将文教授の指導のもとに行われたものです．

主指導教員である萩原将文先生には，学部からの 5 年半，大変お世話になりました．研究内容について示唆に富んだご助言を頂くだけではなく，日頃のご指導の中で，研究の面白さや研究者としての生き方を学ばせて頂きました．また，大変自由な環境で研究をさせていただけたことを改めて深く感謝致します．

ご多忙な中，副査をお引き受け下さり，大変貴重なコメントを下さった斎藤博昭先生，今井倫太先生，篠沢佳久先生に深く感謝申し上げます．

インターンシップの際にお世話になった日本アイ・ビー・エム株式会社東京基礎研究所の皆様に感謝致します．特にメンターの那須川哲也様には，貴重なご助言を数多く頂きました．また，学会などの場で折に触れて気にかけて頂きましたこと，この場をお借りして深く御礼申し上げます．

同じくインターンシップの際にお世話になった株式会社 Preferred Infrastructure, 株式会社 Preferred Networks, 株式会社レトリバの皆様に感謝いたします．特にインターンシップの際にメンター務めて下さいました現株式会社レトリバの西鳥羽二郎様には大変お世話になりました．西鳥羽様には，研究を進めていく上での基礎的な能力となるコーディング力について深く学ばさせて頂きました．西鳥羽様にコードレビューをして頂き，コーディングについて数々のご助言を頂いたことが，研究効率の著しい向上に繋がりました．改めて深く御礼申し上げます．

数多くのご助言をして頂いた萩原研究室の皆様に深く感謝致します．まず学部時代では先輩の本間幸徳さん，菅生健介さん，並木一樹さん，長谷航記さん，沖総一朗さんにお世話になりました．特に本間幸徳さんには卒業後も学会などでお会いした際に，数々のご助言を頂きました．研究室同期の松井辰哉君，和泉清矢君，関悠介君，佛木

真穂さん，小川早紀さんに感謝致します．同期の皆様と過ごした時間は研究内容のみならず，人間的にも得るものが大きかったです．

博士課程を共に過ごした留学生の柯遠志さんは，研究に関する深い議論を行うことができる数少ない相手でした．ディスカッションを通じて得られた数多くの示唆は研究を進める上で大きな指針となりました．ありがとうございます．

後輩の三上佳孝君，迫田真太郎君に感謝致します．三上君がいなければ，私の研究のモチベーションもここまで高く保たれることはなかったと思います．数々の勉強会を開催し，議論を交わした日々は私にとって大切な思い出です．ありがとうございます．迫田君には，本論文の添削をして頂き，貴重なコメントを数多く頂きました．お陰様で本論文をこうして書き切ることができました．お忙しい中，本当にありがとうございました．

自身が博士課程を目指すきっかけを下さった山根宏彰さんに深く感謝致します．研究室を離れた後も度々相談に乗って頂き，自身の研究について有益なアドバイスをいただきました．山根さんがいらっしゃらなければ，本論文は書かれなかったものと思います．本当にありがとうございました．

その他，同じ研究室で過ごした多くの方々に深く感謝致します．この研究室で皆様と切磋琢磨しながら研究できたことをこの上なく嬉しく思っております．この場をお借りして御礼申し上げます．

この場には書ききれませんが，学会や勉強会などで数多くの方々にお世話になりました．特に日本感性工学会，日本知能情報ファジィ学会，言語処理学会，人工知能学会でディスカッションを行って頂いた方々，私の論文の査読をして頂き貴重なコメントを下さった方々に深く感謝致します．また，NLP 東京 D の会では同期の博士課程の学生から数多くの有益なアドバイスを頂きました．本当にありがとうございます．

最後に，自分の博士課程進学についてご理解を頂き，暖かく支えて下さった祖父母，両親，妹，私を明るく励まし続けてくれた妻に深く感謝し，謝辞とさせて頂きます．

付録 A

倫理憲章

A.1　人工知能学会 倫理指針

以下に，人工知能学会の倫理指針について記す．A.2 の「アシロマ AI 23 の原則」と比べると，主に人工知能研究者が持つべき倫理について述べられているのが特徴である．

A.1.1　原文の引用 [1]

1. **人類への貢献**: 人工知能学会会員は，人類の平和，安全，福祉，公共の利益に貢献し，基本的人権と尊厳を守り，文化の多様性を尊重する．人工知能学会会員は人工知能を設計，開発，運用する際には専門家として人類の安全への脅威を排除するように努める．

2. **法規制の遵守**: 人工知能学会会員は専門家として，研究開発に関わる法規制，知的財産，他者との契約や合意を尊重しなければならない．人工知能学会会員は他者の情報や財産の侵害や損失といった危害を加えてはならず，直接的のみならず間接的にも他者に危害を加えるような意図をもって人工知能を利用しない．

3. **他者のプライバシーの尊重**: 人工知能学会会員は，人工知能の利用および開発において，他者のプライバシーを尊重し，関連する法規に則って個人情報の適正な取扱いを行う義務を負う．

4. **公正性**: 人工知能学会会員は，人工知能の開発と利用において常に公正さを持ち，人工知能が人間社会において不公平や格差をもたらす可能性があることを認識

し，開発にあたって差別を行わないよう留意する．人工知能学会会員は人類が公平，平等に人工知能を利用できるように努める．

5. **安全性**: 人工知能学会会員は専門家として，人工知能の安全性及びその制御における責任を認識し，人工知能の開発と利用において常に安全性と制御可能性，必要とされる機密性について留意し，同時に人工知能を利用する者に対し適切な情報提供と注意喚起を行うように努める．

6. **誠実な振る舞い**: 人工知能学会会員は，人工知能が社会へ与える影響が大きいことを認識し，社会に対して誠実に信頼されるように振る舞う．人工知能学会会員は専門家として虚偽や不明瞭な主張を行わず，研究開発を行った人工知能の技術的限界や問題点について科学的に真摯に説明を行う．

7. **社会に対する責任**: 人工知能学会会員は，研究開発を行った人工知能がもたらす結果について検証し，潜在的な危険性については社会に対して警鐘を鳴らさなければならない．人工知能学会会員は意図に反して研究開発が他者に危害を加える用途に利用される可能性があることを認識し，悪用されることを防止する措置を講じるように努める．また，同時に人工知能が悪用されることを発見した者や告発した者が不利益を被るようなことがないように努める．

8. **社会との対話と自己研鑽**: 人工知能学会会員は，人工知能に関する社会的な理解が深まるよう努める．人工知能学会会員は，社会には様々な声があることを理解し，社会から真摯に学び，理解を深め，社会との不断の対話を通じて専門家として人間社会の平和と幸福に貢献することとする．人工知能学会会員は高度な専門家として絶え間ない自己研鑽に努め自己の能力の向上を行うと同時にそれを望む者を支援することとする．

9. **人工知能への倫理遵守の要請**: 人工知能が社会の構成員またはそれに準じるものとなるためには，上に定めた人工知能学会員と同等に倫理指針を遵守できなければならない．

A.2 アシロマ AI 23 の原則

本節では 2017 年 2 月 3 日に発行されたアシロマ AI 23 の原則について述べる．この原則は，2017 年 1 月にカリフォルニア州アシロマにて全世界から集まった専門家達によって議論されたものである．主に,「人類にとって有益な AI とは何か」について議論されており，発表された原則は AI の研究，倫理・価値観，将来的な問題の 3 つの分野に関して言及されている．この原則には 2017 年 9 月の段階で約 1200 人の人工知能研究者の署名が集まっている.

本研究と特に関係が深い項目は，“Ethics and Values” の項である．この項目では,「高度な自律的な人工知能システムは, その目的と振る舞いが確実に人間の価値観と調和すべき」と定められている．この目標を達成するためには, 人間が当たり前のように保持している行動に関する価値観, つまり本研究で対象とするような道徳的常識が必要であると考えられる.

A.2.1 原文の引用 [2]

Research Issues

1. **Research Goal**: The goal of AI research should be to create not undirected intelligence, but beneficial intelligence.

2. **Research Funding**: Investments in AI should be accompanied by funding for research on ensuring its beneficial use, including thorny questions in computer science, economics, law, ethics, and social studies, such as:

    - How can we make future AI systems highly robust, so that they do what we want without malfunctioning or getting hacked?
    - How can we grow our prosperity through automation while maintaining people ’ s resources and purpose?
    - How can we update our legal systems to be more fair and efficient, to keep pace with AI, and to manage the risks associated with AI?

- What set of values should AI be aligned with, and what legal and ethical status should it have?

3. **Science-Policy Link**: There should be constructive and healthy exchange between AI researchers and policy-makers.

4. **Research Culture**: A culture of cooperation, trust, and transparency should be fostered among researchers and developers of AI.

5. **Race Avoidance**: Teams developing AI systems should actively cooperate to avoid corner-cutting on safety standards.

Ethics and Values

1. **Safety**: AI systems should be safe and secure throughout their operational lifetime, and verifiably so where applicable and feasible.

2. **Failure Transparency**: If an AI system causes harm, it should be possible to ascertain why.

3. **Judicial Transparency**: Any involvement by an autonomous system in judicial decision-making should provide a satisfactory explanation auditable by a competent human authority.

4. **Responsibility**: Designers and builders of advanced AI systems are stakeholders in the moral implications of their use, misuse, and actions, with a responsibility and opportunity to shape those implications.

5. **Value Alignment**: Highly autonomous AI systems should be designed so that their goals and behaviors can be assured to align with human values throughout their operation.

6. **Human Values**: AI systems should be designed and operated so as to be compatible with ideals of human dignity, rights, freedoms, and cultural diversity.

7. **Personal Privacy**: People should have the right to access, manage and control the data they generate, given AI systems ' power to analyze and utilize that data.

8. **Liberty and Privacy**: The application of AI to personal data must not unreasonably curtail people ' s real or perceived liberty.

9. **Shared Benefit**: AI technologies should benefit and empower as many people as possible.

10. **Shared Prosperity**: The economic prosperity created by AI should be shared broadly, to benefit all of humanity.

11. **Human Control**: Humans should choose how and whether to delegate decisions to AI systems, to accomplish human-chosen objectives.

12. **Non-subversion**: The power conferred by control of highly advanced AI systems should respect and improve, rather than subvert, the social and civic processes on which the health of society depends.

13. **AI Arms Race**: An arms race in lethal autonomous weapons should be avoided.

Longer-term Issues

1. **Capability Caution**: There being no consensus, we should avoid strong assumptions regarding upper limits on future AI capabilities.

2. **Importance**: Advanced AI could represent a profound change in the history of life on Earth, and should be planned for and managed with commensurate care and resources.

3. **Risks**: Risks posed by AI systems, especially catastrophic or existential risks, must be subject to planning and mitigation efforts commensurate with their expected impact.

4. **Recursive Self-Improvement**: AI systems designed to recursively self-improve or self-replicate in a manner that could lead to rapidly increasing quality or quantity must be subject to strict safety and control measures.

5. **Common Good**: Superintelligence should only be developed in the service of widely shared ethical ideals, and for the benefit of all humanity rather than one state or organization.

付録 B

単語の分散表現の学習方法

本章では，単語の分散表現を学習する手法について説明する．

B.1　単語の分散表現

単語の分散表現とは，各単語を数百から数千次元のベクトルで表現したものである．一般的に単語は離散的な表現であるため，数値的な計算は困難であるが，分散表現を用いることで，単語間の意味の類似性などをモデル化することができる．

B.2　word2vec

B.2.1　概要

単語の分散表現学習の際には，大規模なテキストデータを入力とし，各単語の分散表現を求める．基本的なアイデアは，ある単語の分散表現を用いることで，その周辺の文脈の単語を予測できるように学習を行うというものである．具体的には文 $s(w_1, w_2, ..., w_T)$ があったとき，以下の目的関数を最大化する．

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{\substack{-c\leq j\leq c\\ j\neq 0}}\log P(w_{t+j}|w_t) \tag{B.1}$$

ここで c は前後何単語まで考慮するかを決定するパラメータである．式 (B.1) から，前後文脈は同一に扱い，近傍単語の語順については考慮しないことが分かる．

Skip-gram モデルでは，文書中のある単語 w_I の近傍で単語 w_O が出現する確率は以下で定義される．

$$P(w_o|w_I) = \frac{\exp({v'_{w_O}}^{\mathrm{T}} v_{w_I})}{\sum_{o=1}^{V} \exp({v'_{w_o}}^{\mathrm{T}} v_{w_I})} \tag{B.2}$$

ここで V はコーパス全体での語彙数である．v_w および v'_w は単語 w の分散表現であり，v_w を入力ベクトル，v'_w を出力ベクトルと呼ぶ．入力ベクトルは前後文脈の単語を予測するために使われるベクトルであり，出力ベクトルは，前後文脈の単語ベクトルとして用いられる．

式 (B.2) の分母はベクトル同士の内積および指数の計算を語彙数分繰り返す必要がある．一般的に語彙数は $10^5 \sim 10^6$ のオーダーなのでこの計算は非常に重たい．そこで，すべての計算を行わず，少量のサンプリングされた単語のみに対して計算を行う Negative Sampling という手法が提案されている．

具体的には $\log P(w_O|w_I)$ を以下の式で置き換える．

$$\log P(w_O|w_I) = \log \sigma({v'_{w_O}}^{\mathrm{T}} v_{w_I}) + \sum_{i=1}^{k} \mathbb{E}_{w_i \sim P_n(w)} [\log \sigma({v'_{w_i}}^{\mathrm{T}} v_{w_I})] \tag{B.3}$$

ここで $P_n(w)$ は雑音分布と呼ばれ，単語をサンプリングするための分布である．通常，k は数十程度で十分であり，元の語彙数を考えれば大幅な高速化であると考えられる．

付録 C

ロジスティック回帰モデルを用いた道徳判断

C.1 ロジスティック回帰モデルによる学習

C.1.1 入力

機械学習モデルを学習する際の入力は，訓練データ集合全体である．これらのデータを用いてロジスティック回帰モデルのパラメータの学習が行われる．

C.1.2 特徴量抽出

自然言語文はあくまでもテキスト情報であるため，機械学習モデルが学習可能なようにベクトル情報に変換する必要がある．これを特徴量抽出と呼ぶ．特徴量抽出には自然言語処理で一般的とみなされている単語 N-gram を用いる．

C.1.3 機械学習モデル

機械学習モデルではロジスティック回帰を用いる．ロジスティック回帰モデルは分類モデルである．

入力を $\boldsymbol{x}$ としたとき，出力 y が 1，つまりその入力文が Positive である確率は，

$$P(y = +1 \mid \boldsymbol{x}) = \frac{1}{1 + \exp(-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x})} \tag{C.1}$$

ここで，$\boldsymbol{w}$ は学習されるパラメータベクトルであり，T は転置を表す．

逆に，入力文が Negative である確率は，

$$P(y = -1 \mid \boldsymbol{x}) = 1 - P(y = +1 \mid \boldsymbol{x}) = 1 - \frac{1}{1 + \exp(-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x})} \tag{C.2}$$

上記の数式をまとめると，以下の通りになる．

$$P(y \mid \boldsymbol{x}) = \frac{1}{1 + \exp(-y\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x})} \tag{C.3}$$

ここで，y はラベル情報を表し，Positive であれば $+1$ をとり，Negative であれば -1 を取る．学習の際には，以下の目的関数を最小化するようにパラメータベクトル $\boldsymbol{w}$ の更新が行われる．

$$L(\boldsymbol{w}) = -\sum_{i=1}^{N} \log p(y^{(i)} \mid \boldsymbol{x}^{(i)}) + c_1 r_1(\boldsymbol{w}) \tag{C.4}$$

ここで，c_1 はハイパーパラメータであり，N は訓練データの数を表す．$r_1(\boldsymbol{w})$ は L2 正則化項であり，以下の通りに定義される．

$$r_1(\boldsymbol{w}) = \|\boldsymbol{w}\|^2 \tag{C.5}$$

C.1.4 出力

機械学習モデルの出力は，学習済みのモデルである．このモデルを用いて評価用の入力文に対して道徳判断が行われる．

C.2 学習済みモデルを用いた道徳判断

C.2.1 入力

道徳判断タスクは 2 章と同様であり，1 文を入力とする．入力された文はまず形態素解析器 MeCab を用いて分かち書きが行われる．その後，上記で説明した方法により，実次元のベクトル表現に変換が行われる．最後に変換されたベクトル表現及び，C.1.3 で学習されたモデルを用いて道徳判断が行われる．

C.2.2 出力

具体的には，以下の通りに判断が行われる．

$$Output = \begin{cases} Positive & (P(y = +1 \mid \boldsymbol{x}) > 0.5) \\ Negative & (otherwise) \end{cases} \tag{C.6}$$